Livre blanc sur les grilles de production

- Institut Des Grilles

Livre Blanc
sur les grilles de production

Intérêt scientifique et utilisation

Bilan, perspectives et recommandations

mai 2009

TABLE DES MATIÈRES

RÉSUMÉ

Depuis 10 ans, les grilles informatiques font l'objet d'une activité de recherche et développement intensive en Europe. La communauté de physique des particules a joué un rôle pionnier et moteur autour du CERN qui lui permet de disposer aujourd'hui d'une infrastructure de production opérationnelle pour l'analyse des données du LHC. D'autres communautés de recherche, notamment en sciences du vivant, de la planète et de l'univers se sont intéressées très tôt à l'utilisation des grilles informatiques et ont démontré leur intérêt pour le traitement de grands volumes de données distribuées.

Ce livre blanc s'inscrit dans le contexte d'un exercice de prospective nationale sur l'intérêt scientifique des grilles de production. Les quatre questions sous-jacentes étaient les suivantes :

- Les grilles de production constituent-elles une technologie nouvelle capable d'apporter aux chercheurs des différentes disciplines un outil décisif pour réaliser des percées scientifiques ? Dans quels domaines plus particulièrement ? Quels succès ont été obtenus et quels sont les objectifs à court et moyen terme de chaque communauté ?

- Quelle place doivent occuper les grilles de production dans l'«écosystème » des moyens de calcul offerts à la communauté de recherche française ?

- Quel investissement matériel est nécessaire pour répondre aux demandes émanant des différentes communautés ?

- Quels sont les besoins humains nécessaires et de quel type ?

Il ressort de la prospective que les grilles de production sont devenues en quelques années un outil indispensable à la communauté nationale dans plusieurs domaines scientifiques très importants : physique subatomique, sciences du vivant, sciences de la planète principalement. Les grilles de production apparaissent très clairement comme la ressource informatique, complémentaire des grands supercalculateurs, qu'il faut mettre à la disposition du plus grand nombre. Un grand effort de formation et d'information reste à accomplir car tous les domaines ne sont pas encore impliqués au même niveau et, dans chaque domaine, tous les chercheurs ne sont pas informés ou formés au même niveau.

Les grilles de production doivent donc occuper une place reconnue et bien documentée au plus haut niveau dans l'écosystème des moyens de calcul.

Une réponse précise concernant les besoins matériels n'a été fournie que par la communauté de physique subatomique. Celle-ci représente aujourd'hui 2/3 de l'utilisation totale des grilles mais ce chiffre pourrait descendre à 50% si les autres communautés amplifient leur utilisation. Le besoin total peut donc être estimé entre 1,5 et 2 fois les besoins exprimés par la physique subatomique, soit un besoin total de 100 kSI2k[1] et 75 PétaOctets d'ici 2012, soit un besoin de financement de 38,5 M€ qui pourra être satisfait par les sommes inscrites au budget TGE pour le Tier-1 à Lyon (environ 20 M€ sur la période considérée) et par un plan de financement de 18,5 M€ sur 5 ans soumis au Comité de Pilotage national des Grilles de production.

Ces réponses ressortent des recommandations de huit groupes de travail qui sont documentées dans le livre blanc. Cinq groupes de travail thématiques se sont réunis au cours de l'année 2008 pour étudier l'utilisation des grilles en biologie et santé, chimie, sciences humaines et sociales, sciences de l'ingénieur et informatique et sciences de la planète et de l'univers. Chaque groupe a fait circuler un sondage dans sa communauté : ces sondages ont été dans l'ensemble bien accueillis et plus de 1000 réponses ont été reçues et analysées. Les groupes ont mis en avant des besoins similaires en mettant chacun l'accent sur des priorités spécifiques :

- le groupe de travail biologie-santé a souligné la nécessité du recrutement d'ingénieurs qui servent d'intermédiaire entre les utilisateurs finaux et les équipes qui conçoivent les middlewares et administrent les sites de production.

- Le groupe de travail planète-univers a mis l'accent sur la nécessité que la grille s'intègre aux côtés des supercalculateurs à l'écosystème spécifique en supportant les environnements de travail les plus courants et en s'interfaçant efficacement aux centres de données propres à la communauté

- Le groupe de travail Physique des Particules a insisté sur la montée en puissance de l'utilisation de la grille dans la communauté avec le démarrage du LHC

- Le groupe de travail Sciences de l'Ingénieur et Informatique a souligné l'importance de développer des passerelles vers la communauté de recherche sur les grilles et la nécessité de mettre à disposition des logiciels commerciaux sur la grille de production

- Le groupe de travail chimie a souligné l'importance de créer un noyau dur d'utilisateurs de la grille de production en chimie pour accroître son adoption

- Le groupe de travail Sciences Humaines et Sociales a souligné le rôle structurant de la grille pour la communauté et son intérêt pour renforcer l'activité des Centres de Ressources Numériques

À côté des groupes thématiques, trois groupes transverses ont étudié respectivement les enjeux liés aux grilles de données, aux grilles régionales et à la relation entre grilles et supercalculateurs. Les travaux de ces groupes ont permis de faire un certain nombre de constats et ont débouché sur quelques recommandations intéressant toutes les communautés scientifiques :

- le rôle très important des grilles régionales pour identifier et intégrer de nouveaux utilisateurs des grilles montre l'importance d'accroître le nombre de sites de la grille de production nationale. Il est particulièrement important d'implanter ces nouveaux sites dans les grandes villes universitaires comme Bordeaux, Lille, Nancy, Rennes ou Toulouse

- l'importance du rapprochement entre la communauté développant la grille de production et les équipes des centres de calculs équipés de supercalculateurs car les ressources proposées sont totalement complémentaires

- la complexité des données, la croissance rapide voire exponentielle dans certaines communautés de leur volume et la multiplication des outils de gestion distribuée montrent l'importance de poursuivre la réflexion du groupe de travail sur les grilles de données

En conclusion, le déploiement d'une grille pluridisciplinaire de production en France constitue une étape majeure dans l'intégration de la France dans l'Espace Européen de la Recherche et sera dans les années à venir un support essentiel pour les chercheurs français dans la compétition internationale.

[1] Ksi2k = 1000 SpecInt 2000. Les cœurs de processeur utilisés couramment ont une puissance variant entre 1,5 et 2,5 ksi2k suivant leur nature et leur fréquence d'horloge.

EXECUTIVE SUMMARY

For 10 years, computing grids have been the object of intense research and development in Europe. The particle physics community has played a pioneering and leading role around CERN : as a consequence, it can benefit today of an operational production grid infrastructure for the analysis of LHC data. Other research communities, particularly in the field of life, planet and universe sciences, have shown early interest for computing grids for the treatment of large volumes of distributed data.

This white paper has been written in the context of a national prospective about the scientific interest of production grid infrastructures. The four underlying questions were the following :

- Are production grids key tools to enable front-end scientific research in many disciplines ? Which scientific domains would be most impacted ? Which successes have been achieved so far and what are the short and mid term objectives of the communities using the grid ?
- Which could be the role of production grids within the “ecosystem” of resources offered to the French research community ?
- Which hardware investment is needed to meet the needs of the different communities ?
- What are the human resources needed expressed both in terms of volume and expertise ?

Production grids have become in a few years a mandatory tool for the national research communities in several disciplines, especially subatomic physics, life sciences and planet sciences. Production grids appear very clearly as providing computing resources complementary to those offered by supercomputers. A large effort is still needed in terms of information and training as all scientific disciplines have not been exposed at the same level to this new technology.

Production grids must therefore have a recognized and documented role in the ecosystem of computing resources in France.

A precise answer on the amount of equipment needed was only provided by subatomic physics which represents today about 2/3 of the total usage of grids in France and in Europe but this fraction should decrease to about 50% as other disciplines increase their activity. A reasonable estimate of the overall hardware needed can therefore be estimated to about 1.5 to 2 times the requirements expressed by the subatomic physics community, corresponding to a total need for 100kSI2k and 75 PetaOctets by 2012. The corresponding budget is approximately 38.5 M€ which is covered by the TGE budget for Lyon LHC Tier-1 Computing Centre (about 20M€ during the period considered) and by a 18.5 M€ budget plan over 5 years submitted to the national Production Grid Steering Committee.

These answers come out of the recommendations of 8 working groups documented in the white paper. Five thematic groups met during year 2008 to study the use of production grids in medical and life sciences, chemistry, human and social sciences, engineering and computer sciences, planet and universe sciences. For this purpose, each working group surveyed the community and more than 1000 answers were collected and analysed. The groups expressed very similar requirements with different highlights:

- The life sciences group stressed the need to hire engineers who would act as mediators between end users and middle ware developers and site administrators.
- The planet-universe working group stressed the importance for the production grid to get integrated besides supercomputers into the present ecosystem of computing resources, insisting on the necessity to have the right interfaces to the existing working environments and data centres used by the community.
- The particle physics working group insisted on the necessary scaling up of the production grid resources with LHC kick-off.
- The Computer and Engineering sciences working group highlighted the need to build bridges with the research community on grids and the necessity to deploy licensed software on production grids.
- The chemistry working group stressed the need to create a hard core of users of the production grid to increase its adoption.

- The human and social sciences working group highlighted the potential structuring role of the production grid for its community and its interest to strengthen the activity of Digital Resource Centres (Centres de Ressources Numériques).

Besides thematic groups, three transverse groups studied specific issues related respectively to data grids, regional grids and the relationship between grids and supercomputers. Their activity allowed to identify a number of issues and to propose recommendations relevant to all research communities:

- The specific role of regional grids to identify and integrate new users has been acknowledged. As a consequence, it was recommended to install new production grid nodes in the major French university cities where they are missing today, especially, Bordeaux, Lille, Nancy, Rennes and Toulouse.

- The need has been stressed to develop collaboration between the operational teams of the supercomputing centres and the production grid nodes as the resources proposed to the research communities are fully complementary.

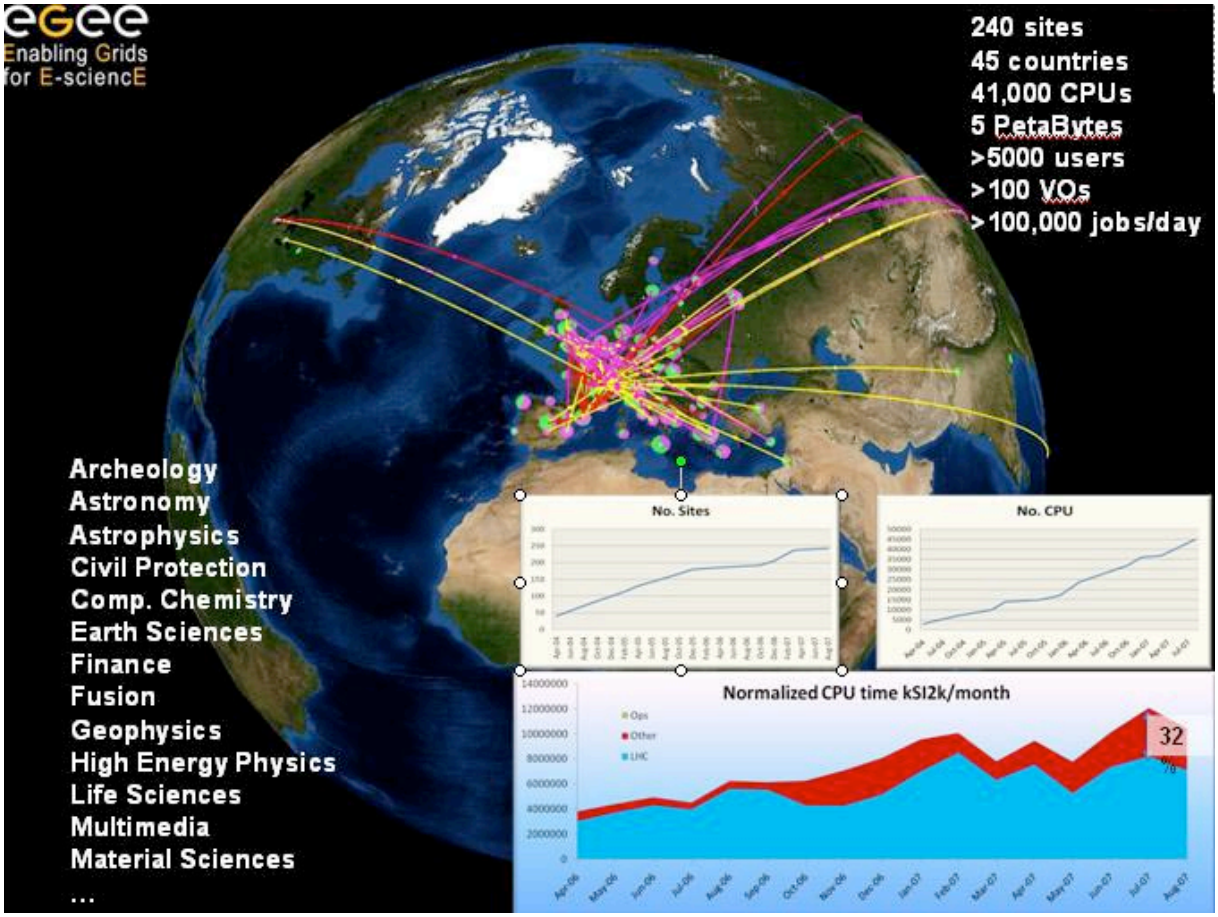
- In view of the growing volume and complexity of the data to be analysed across the scientific disciplines, the need for pursuing a reflection on the best strategy for data integration and inter operability has been expressed.

In conclusion, the deployment of a multidisciplinary production grid in France is a major step towards the integration of French research communities in the European Research Area and will certainly be a major competitive advantage for French scientists in the coming years.

INTRODUCTION

Ce Livre Blanc a pour ambition de faire le point sur l'intérêt scientifique des grilles de production et leur degré d'utilisation présent ou futur dans les différents domaines scientifiques. Il s'agit de mesurer l'importance que cet outil doit prendre dans l'écosystème informatique mis à la disposition des chercheurs pour leur permettre de réaliser les percées scientifiques rendues possibles par l'ensemble de ces nouvelles technologies.

Il s'inscrit dans un contexte en forte évolution. En effet, la communauté de physique des particules a atteint l'objectif très ambitieux qu'elle s'était fixée en 2000, à savoir de disposer d'une infrastructure de grilles de production de très grande envergure stable et efficace, permettant l'exploitation complète des données issues du collisionneur LHC. La grille de production a montré tout au long de l'année 2008 qu'elle était fin prête à recevoir les données que le LHC aurait dû délivrer à l'automne 2008. Depuis 2004 et le programme européen EGEE, cette grille de production est également ouverte à l'ensemble des autres disciplines scientifiques qui mobilisent de façon constante environ un tiers des ressources totales de la grille (80000 unités de calcul et 15 Péta octets de stockage répartis sur plus de 250 sites à travers le monde, voir figure 1).



L'avènement de la grille de production comme un outil de référence pour de très grandes infrastructures de recherche comme le LHC aux très grandes constantes de temps, nécessite une pérennisation tant à l'échelle européenne que nationale. Le document « EGI Blueprint » jette les bases de ce que devra être la structure d'une telle organisation pérenne, basée sur des entités fortes au niveau de chaque pays appelées NGI (National Grid Initiatives) en charge de la coordination au plus niveau de tous les aspects de la grille nationale et d'une entité légale européenne EGI.org chargée de la coordination de la grille européenne et de la fourniture des fonctions centrales nécessaires à son fonctionnement.

L'exercice de prospective décrit dans ce livre blanc découle directement de ces initiatives puisque l'ensemble de la démarche décrite ci-dessous est régi par le protocole d'accord national signé en automne 2008 entre le Ministère de la recherche, le CNRS, le CEA, l'INRIA, l'INRA, RENATER et la conférence des Présidents d'Universités. Ce protocole préfigure la NGI France en cours de création autour de l'Institut des Grilles du CNRS, créé le 1er septembre 2007 pour fédérer l'ensemble des activités relatives aux grilles à l'intérieur du CNRS. Ce protocole d'accord institue en particulier un Comité de Pilotage National, commanditaire de ce Livre Blanc et qui a confié de façon alors informelle l'organisation de la prospective nationale à l'Institut des Grilles en janvier 2008. Il est important de souligner que le décalage de neuf mois environ entre le lancement de l'initiative de prospective et la signature formelle du Protocole d'Accord National n'a finalement que peu

perturbé la prospective grâce à la très bonne coopération qui a régné entre tous les acteurs quels que soient leurs organismes d'origine. À titre d'exemple, plusieurs chercheurs de l'INSERM ont joué un rôle d'animation important même si cet organisme n'a pas encore signé le protocole d'accord.

La démarche prospective

La démarche prospective dont ce livre est l'aboutissement a donc démarré début 2008 par la création de 8 groupes de travail thématiques et de 6 groupes transversaux dont la liste est indiquée ci-dessous :

- Thématiques

– Biologie santé
– Sciences de la Planète-Sciences de l'Univers
– Sciences Humaines
– Chimie
– Sciences de l'Ingénieur et Informatique
– Physique et Mathématiques
– Physique Subatomique
– Agronomie-Ecologie

- Transverses

– Grilles de données
– Grilles régionales; relation avec GRID5000
– Relation supercalculateurs
– TGE/ESFRI
– Accès à la grille
– Relations avec les Industriels

La mission assignée aux groupes thématiques était triple :

• Faire un bilan de l'utilisation actuelle de la grille de production dans leur domaine respectif en essayant d'en dégager les avantages spécifiques.

• Faire un sondage le plus large possible bien au-delà des utilisateurs actuels pour connaître les futurs besoins et les points bloquants éventuels.

• Dresser une liste de recommandations et de conclusions liées aux deux points précédents.

Chaque groupe thématique a eu toute liberté pour créer un réseau d'animateurs d'horizons aussi variés que possible. Leur liste est donnée dans chaque rapport thématique. On ne peut que se réjouir du grand nombre de personnes impliquées (plus d'une centaine) et surtout du grand nombre de réponses obtenues dans les questionnaires (plus de 3000 chercheurs ayant répondu directement ou indirectement par le biais de leur chef d'équipe). Les résultats obtenus bien que comportant inévitablement un biais d'échantillon sont à notre avis très significatifs.

Il faut noter que sur les 8 groupes thématiques, un rapport n'est pas disponible et un autre n'est que très partiel. En effet, il n'a pas été possible dans le temps imparti d'identifier les bons acteurs dans le domaine de l'agronomie et de l'écologie. Ce domaine est fort peu utilisateur actuel des grilles de calcul mais les premiers contacts laissent à penser que certaines applications pourraient en bénéficier dans un futur proche. Une situation analogue a été rencontrée dans le groupe « Physique-Mathématique » où, en dehors de l'application de fusion nucléaire, les contacts nécessaires n'ont pas encore pu être établis. Dans ces deux secteurs, la prospective devra donc se dérouler en 2009 pour compléter les résultats indiqués dans ce livre.

Les six groupes transverses avaient une mission un peu différente car plutôt que de sonder la communauté au sens large, il s'agissait de réunir un petit nombre d'experts pour essayer de dégager un certain nombre de tendances fortes sur des sujets communs à l'ensemble des disciplines. Le groupe « relations avec les industriels » avait prévu une rencontre en 2008 avec les responsables informatiques des grandes sociétés françaises sur ce thème qui ne pourra avoir lieu qu'en 2009.

Christian Barillot, Christophe Blanchet, Hugues Benoit-Cattin , Vincent Breton, Sorina Camarasu, François Cambien, Christophe Caron, Olivier Collin, Antoine De Daruvar, Frédéric Desprez, Gaël Even, Géraldine Fettahi, Richard Lavery, Hugues Leroy, Tiphaine Martin, Michel Masella, Claudine Médigue, Alexis Michon, Johan Montagnat, Angel Osorio, Frédéric Plewniak, Didier Rognan, Bruno Spataro, El Ghazali Talbi, Thierry Toursel

1. BILAN DE L'INTÉRÊT SCIENTIFIQUE

Les grilles constituent un environnement privilégié pour l'intégration des données et l'accès aux ressources (calculs, stockage) en sciences du vivant et en médecine. En effet, les sciences du vivant sont confrontées à une croissance exponentielle du volume des données produites notamment par la biologie moléculaire. Ces données sont produites dans presque tous les laboratoires et seule une infrastructure distribuée est à même de répondre aux besoins de les traiter, les analyser, les corriger ou les mettre à jour. Dans le domaine de la santé, l'explosion de l'imagerie médicale dans les hôpitaux en France et en Europe et la décision politique de développer le Dossier Médical Partagé placent les grilles au cœur des enjeux de l'infrastructure de santé en France. Mais si leur utilisation pour la clinique est conditionnée par l'évolution des textes de loi, il n'en est pas de même pour la recherche médicale de telle sorte qu'elles sont utilisées aujourd'hui dans plusieurs champs disciplinaires. Dans la partie 1.1, nous illustrerons l'impact des grilles aujourd'hui à travers l'exemple de plusieurs applications scientifiques déployées aujourd'hui en France sur des infrastructures de production.

Pour mieux comprendre les besoins de la communauté et l'intérêt scientifique de la communauté pour les grilles, un sondage a été organisé au printemps 2008 dont les résultats sont discutés dans la partie 1.2.

1.1 Exemples marquants d'utilisation

La grille Décrypthon

Le projet Décrypthon (http://www.decrypthon.fr), lancé en 2004, est le fruit d'une collaboration entre le CNRS, l'Association Française contre les Myopathies (AFM) et la société IBM. Il vise à mettre à la disposition des équipes de recherche en bioinformatique des ressources informatiques de calcul et de stockage. Ces ressources placées dans les universités constituent une plate-forme de type grille comprenant six sites (Bordeaux I, Jussieu, Lille I, ENS Lyon, Pierre et Marie-Curie Orsay, le Crihan à Rouen) reliés par le réseau RENATER. Le Décrypthon met en œuvre les moyens nécessaires à l'exploitation de la grille; il finance des équipes de recherche sélectionnées sur appels d'offres et les accompagne pour les aspects informatiques des projets (modélisation, portage des applications, gestion des données, ...). L'équipe projet GRAAL (CNRS, ENS Lyon, Inria) intervient depuis le début du projet comme expert en «gridification d'applications». En 2006, l'intergiciel DIET (http://graal.ens-lyon.fr/DIET/) développé dans l'équipe-projet GRAAL a été retenu pour assurer la continuité du support de la grille Décrypthon. Il assure la répartition transparente du travail sur l'ensemble des 6 centres de calcul universitaires au travers du réseau RENATER. Par ailleurs, certaines applications utilisent une grille d'Internautes (World Community Grid, WCG, http://www.worldcommunitygrid.org).

La grille universitaire Décrypthon s'articule autour de plusieurs éléments séparés : l'intergiciel de grille DIET, un portail web pour accéder aux ressources de la grille, des gestionnaires locaux de ressources propres à chaque centre de calcul. Le portail Web se trouve sur une machine dédiée à Orsay. Il contient une application web spécifique à chaque projet scientifique leur permettant la soumission de travaux scientifiques sur l'ensemble des ressources Décrypthon. Le portail s'appuie ensuite sur l'intergiciel DIET pour soumettre les calculs sur les serveurs adaptées aux besoins de chaque application.

La plate-forme DIET déployée pour la grille Décrypthon se compose d'un Master Agent hébergé sur une machine à Orsay, et d'un ServerDeamon (SeD) lancé sur chaque frontale d'accès aux ressources des centres de calcul universitaire. Les SeD sont connectés au Master Agent. Ils sont destinés à collecter les informations de disponibilité et de performances des serveurs et à soumettre les jobs aux gestionnaires locaux (Loadleveler, PBS, OAR ...). Le gestionnaire local de ressources (Batch scheduler) est le système propre à chaque centre universitaire qui assure au niveau local la répartition et la bon déroulement des calculs sur les machines scientifiques dédiées. Chaque site définit une politique d'utilisation de leur ressources, les machines dédiées Décrypthon s'intègrent dans le parc des machines scientifiques de l'université. De cette manière l'ensemble des machines d'un centre de calcul sont partagées par les utilisateurs locaux et les utilisateurs des projets Décrypthon. Enfin,

l'intergiciel DIET assure la migration des données nécessaires pour les calculs et le stockage des résultats vers les serveurs dédiés à cet usage. Aujourd'hui, le programme Décrypthon s'ouvre à l'Institut des Grilles dans le cadre d'un partenariat en train d'être mis en place.

NeuroLOG, une grille pour les neurosciences

Les neurosciences computationnelles engagent des études anatomico-fonctionnelles du cerveau sur des cohortes de sujets de tailles importantes, sélectionnées pour leurs caractéristiques propres à l'étude (groupes d'âges, pathologies, etc). L'analyse des images cérébrales joue une rôle croissant dans ce type d'étude. Des chaînes complètes d'analyse d'images sont mises en place et réitérées sur tous les patients d'une étude pour l'extraction de paramètres quantitatifs et statistiques caractérisant les populations considérées. La grille de calcul fournit une infrastructure adaptée à la fédération de données spécifiques, parfois rares, acquises dans différents centres neuro-radiologiques et à l'obtention des calculs résultants. Des travaux sont entrepris, dans le cadre du projet NeuroLOG en particulier, pour assurer l'utilisabilité des grilles (environnements ouverts sur l'Internet à grande échelle) dans le contexte des neurosciences. Ils prennent en compte les contraintes de confidentialités liées à la manipulation de données médicales et les besoin de contrôle d'accès aux ressources. Ils permettent ainsi la conduite d'études multi-centriques sans déroger aux politiques locales de gestion des ressources. En outre, des outils de haut niveau, adaptés à l'usage des neuro-scientifiques sont développés: représentation abstraites des données et interfaces de médiation pour la fédération d'entrepôts hétérogènes; description des chaînes de traitement et interface avec la grille; intégration des résultats d'analyse dans la plate-forme fédérée. Des études sont conduites dans le cadre de trois pathologies : la sclérose en plaques, les accidents vasculaires cérébraux et les tumeurs cérébrales.

WISDOM, recherche de nouveaux médicaments sur la grille

La recherche de nouveaux médicaments est longue et coûteuse : environ 10 ans et 1 milliard de dollars sont nécessaires pour la découverte et le développement de nouveaux médicaments. Le développement de stratégies in silico est l'une des pistes privilégiées pour réduire à la fois le coût et la durée du processus. Parmi ces stratégies, le criblage virtuel constitue l'une des voies les plus prometteuses pour accélérer le choix de molécules potentiellement actives sur une cible biologique donnée : il s'agit de calculer à partir des structures tridimensionnelles de la molécule biologique ciblée et de composés choisis dans une base de données commerciale l'énergie potentielle de liaison à l'aide de logiciels d'ancrage (docking) et de dynamique moléculaire. Cette approche permet de concentrer les tests biologiques sur les composés chimiques les plus prometteurs, ceux dont on attend la plus grande efficacité. Elle est appliquée avec succès depuis 2005 par un consortium international piloté par un laboratoire français pour rechercher de nouveaux médicaments contre le paludisme et la grippe aviaire. En quelques semaines, plusieurs siècles de temps de calcul ont ainsi été utilisés sur quelques milliers de processeurs pour produire des listes courtes et ciblées de molécules prometteuses dont le test in vitro a confirmé l'activité et dont certaines ont fait l'objet d'un dépôt de brevet.

Référence : N. Jacq, J. Salzemann, Y. Legré, M. Reichstadt, F. Jacq, E. Medernach, M. Zimmermann, A. Maaß, M. Sridhar, K. Vinod-Kusam, J. Montagnat, H. Schwichtenberg, M. Hofmann, V. Breton, Grid enabled virtual screening against malaria, Journal of Grid Computing Vol 6 n°1, 29-43, 2008.

1.2 Résultats du sondage

Au printemps 2008, un sondage a été organisé pour mieux comprendre l'intérêt et le niveau d'exposition de la communauté pour les grilles et plus généralement ses besoins. Le sondage très court portait une douzaine de questions posées via internet (http://www.surveymonkey.com). Plus de 400 réponses ont été collectées en quelques semaines, issues d'au moins 60 laboratoires dans 24 villes de France. À la représentativité géographique s'ajoute une représentativité en termes d'organismes et de disciplines :

- les réponses sont venues d'universitaires, de chercheurs permanents (CNRS, INSERM, INRA), de professionnels de santé (médecins, physiciens médicaux) mais aussi de chercheurs non-permanents (doctorants et post-doctorants)

- les personnes interrogées ont été classées en trois domaines thématiques : sciences du vivant, santé et chimio-informatique. La couverture thématique (figure 1 et figure 2) est large mais déséquilibrée entre les sciences du vivant (298 réponses), la recherche médicale (120 réponses) et la chimio-informatique (19 réponses)

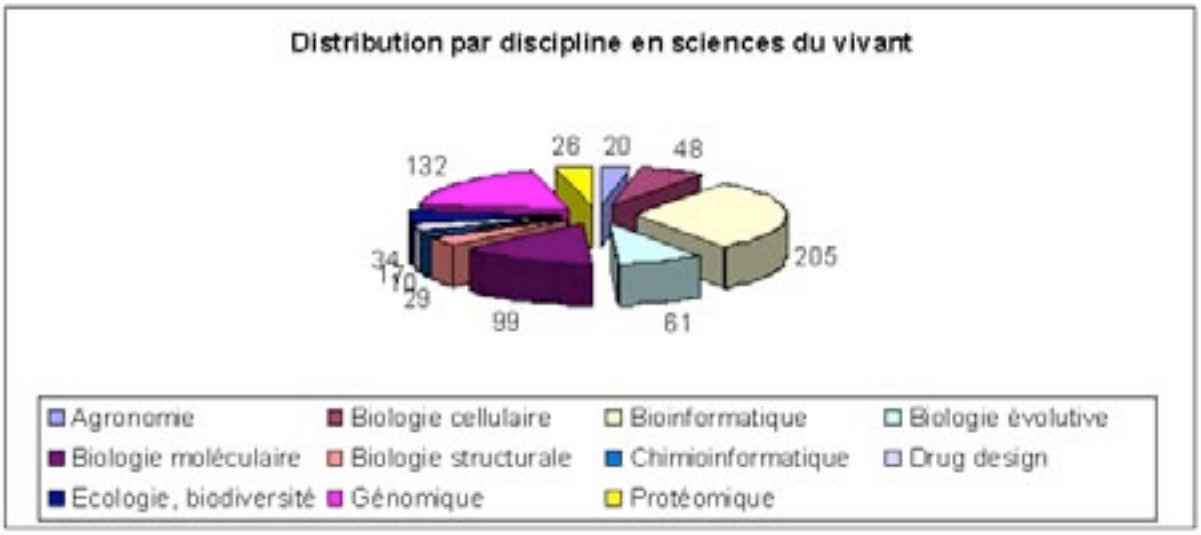


Figure 1 : distribution thématique en sciences du vivant pour 298 personnes sondées (réponses multiples autorisées)

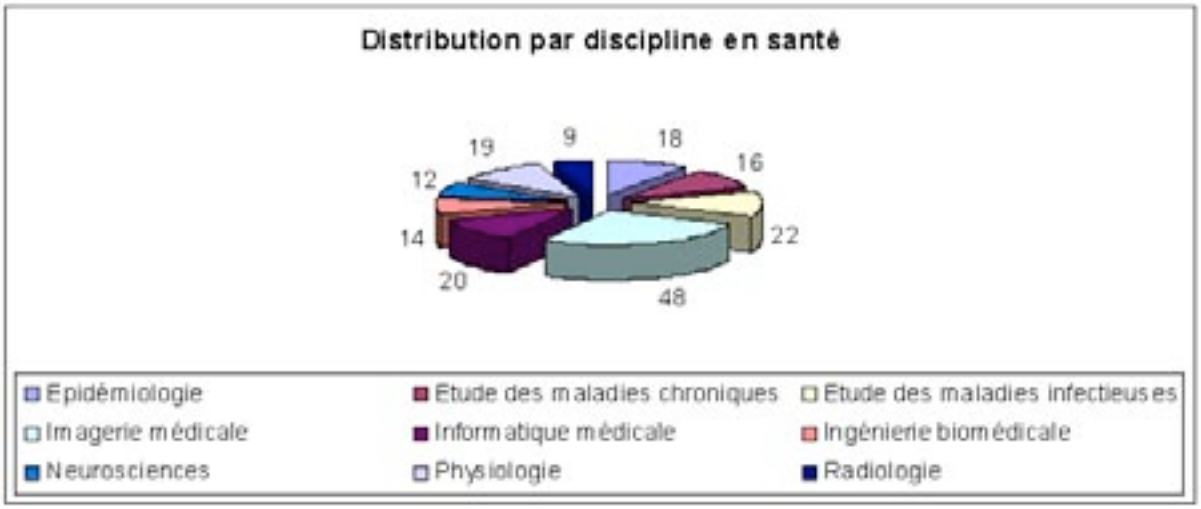


Figure 2 : distribution thématique en santé pour 120 réponses (réponses multiples autorisées)

1.2.1. Connaissance et utilisation des grilles dans la communauté

La figure 3 présente les réponses des personnes sondées à deux questions concernant leur connaissance personnelle des grilles et l'utilisation de celles-ci dans leur laboratoire. Il ressort des réponses qu'une fraction significative de la communauté est informée de l'existence des grilles et qu'elles sont déjà utilisées dans de nombreux laboratoires.

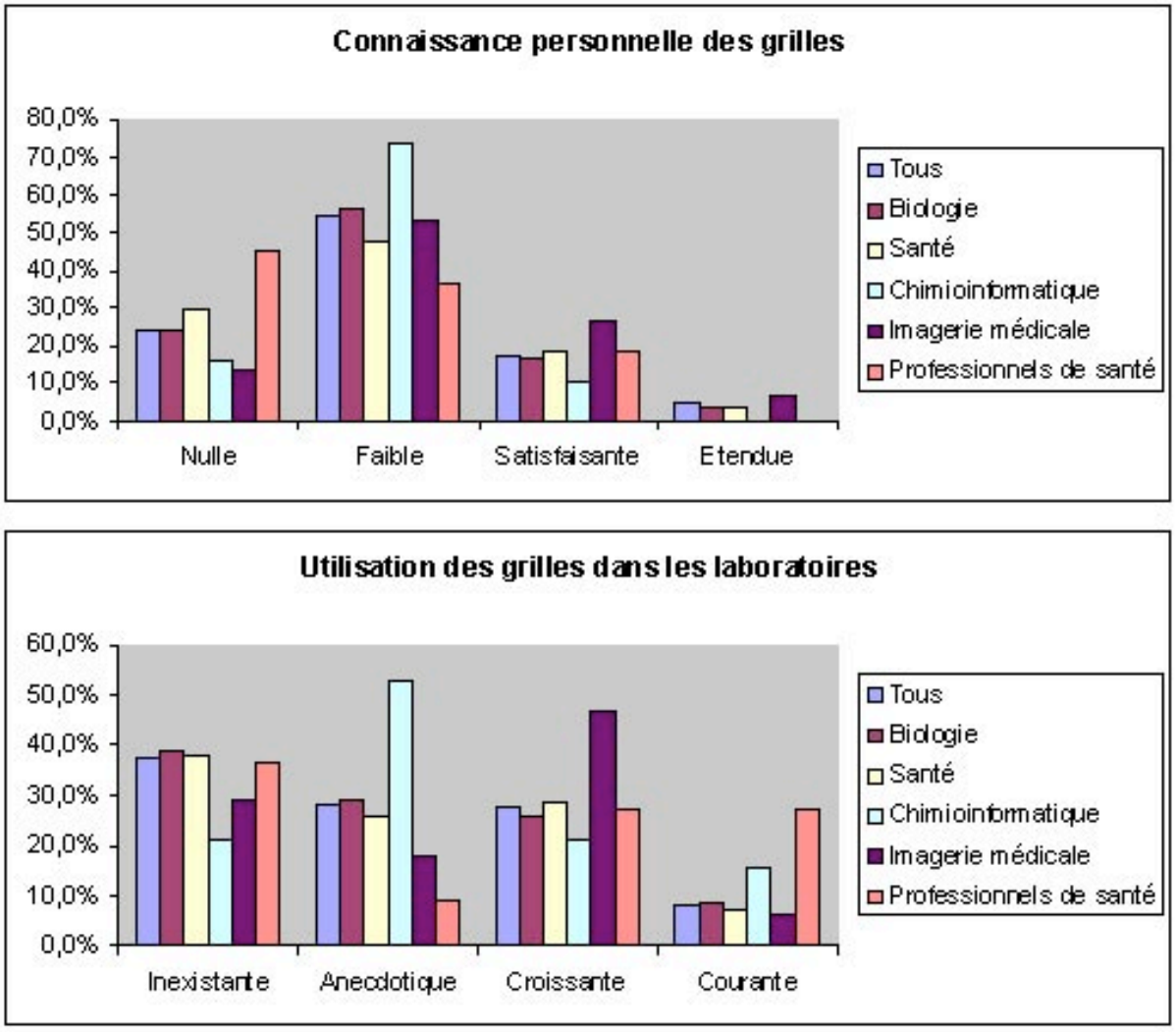


Figure 3 : connaissance et utilisation des grilles dans la communauté

1.2.2 Besoins sur grilles et sur supercalculateurs

La figure 4 propose une comparaison des besoins sur grille et sur supercalculateur. Il est très intéressant de noter que des besoins importants sont exprimés pour les deux architectures. Ce résultat confirme l'importance de fournir à la communauté des ressources de calculs très hétérogènes car les besoins sont très variés.

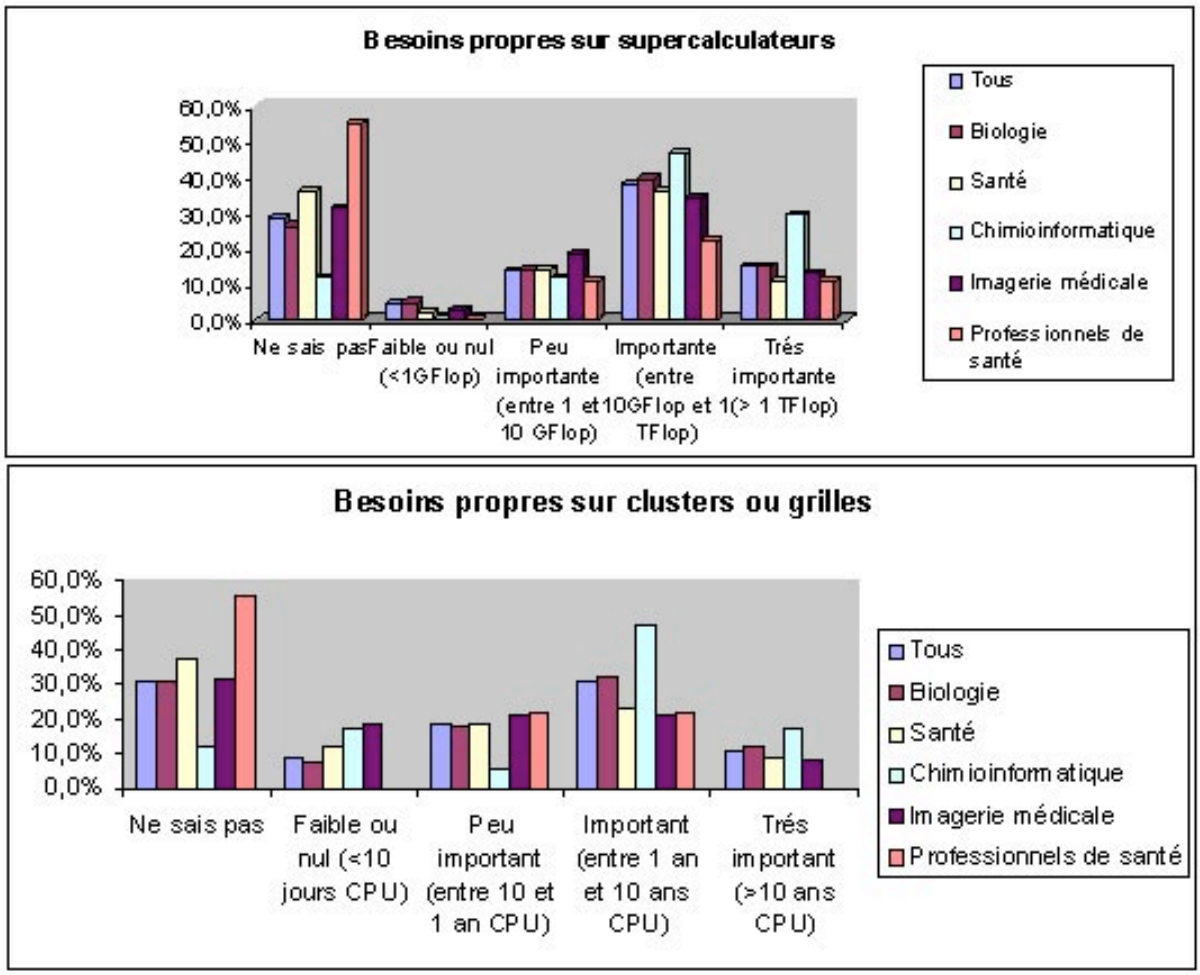


Figure 4 : besoins propres en ressources informatiques

1.2.3 Planification des besoins

La figure 5 présente les besoins en termes de planification des ressources de calculs et de stockage. Elle fait clairement apparaître la nécessité d'accéder à des ressources de calculs à la demande tandis que les besoins de stockage sont beaucoup plus aisés à planifier.

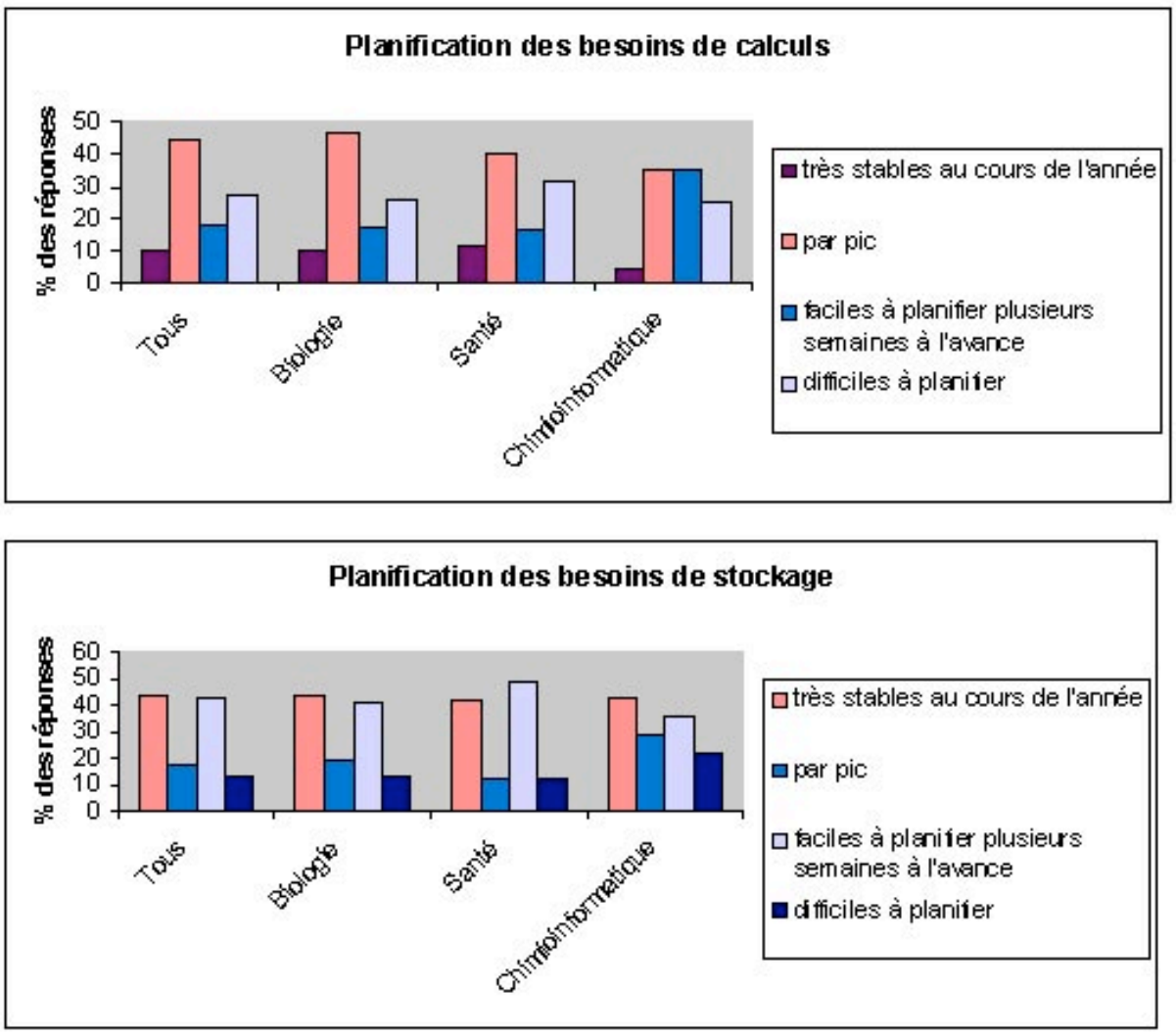


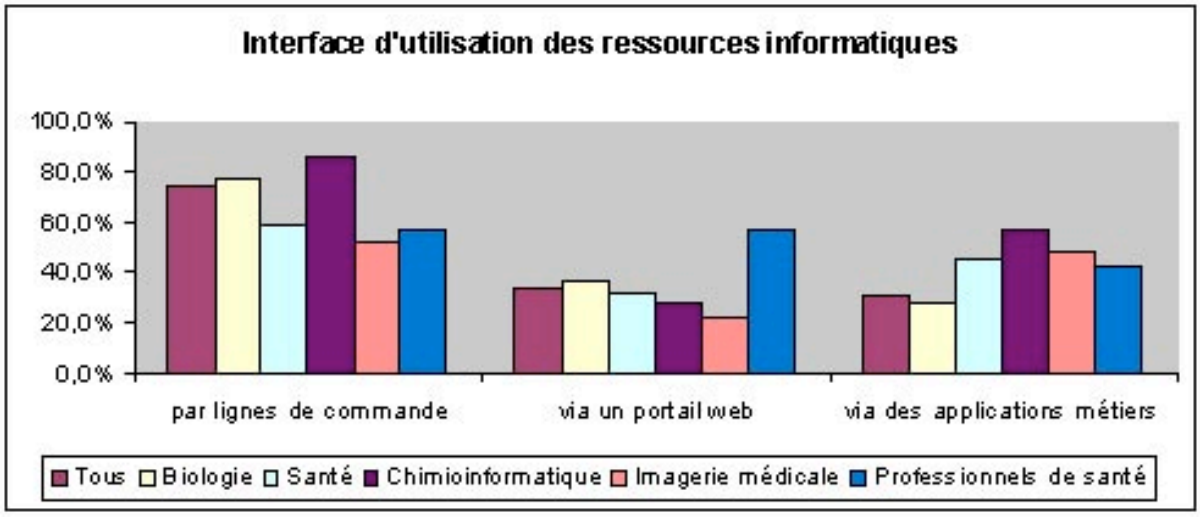
Figure 5 : planification des besoins

1.2.4 Interfaces et sécurité

Comme le montre la figure 6, la communauté souhaite pouvoir accéder aux ressources par lignes de commande, à travers des interfaces web et via des applications métiers.

En ce qui concerne la sécurité, les communautés en biologie et en santé expriment des besoins significativement différents :

- la communauté de biologie se satisfait très majoritairement d'une architecture de sécurité limitée au contrôle d'accès
- par contre, il apparaît clairement que la capacité à encrypter et anonymiser les données est indispensable pour les applications médicales sur les grilles.



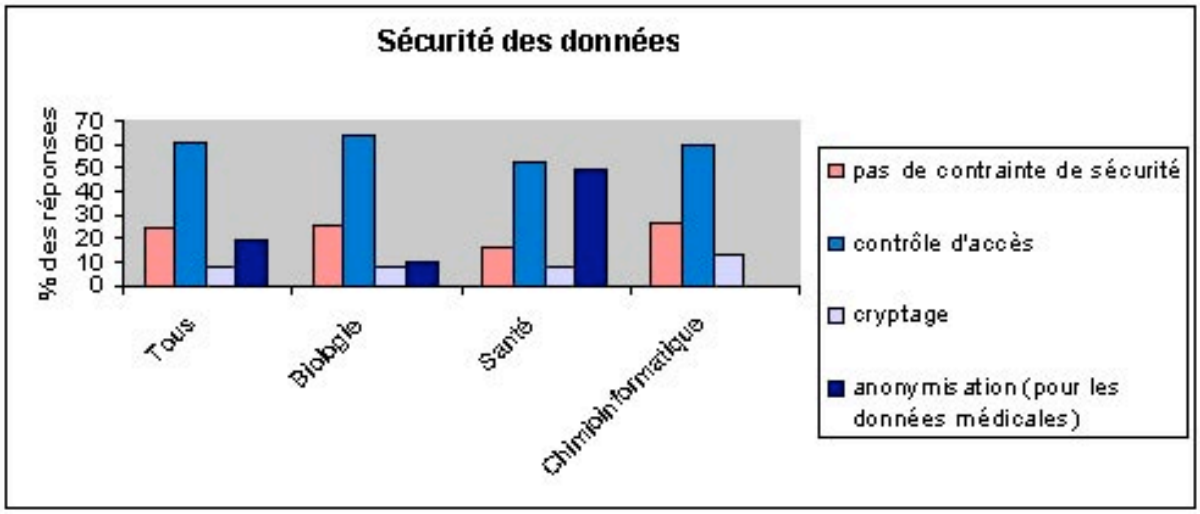


Figure 6 : interfaces et sécurité

1.2.5 Conclusion

Le sondage fait clairement apparaître l'intérêt de la communauté dans toute sa variété pour les grilles. Du médecin au chimiste en passant par le biologiste, tous expriment un besoin pour des ressources de calcul et de stockage distribués sur une grille. Ce besoin n'est pas exclusif et un besoin complémentaire est exprimé pour des ressources centralisées sur supercalculateur. Il est à noter la très grande homogénéité des réponses au travers de la communauté sur toutes les questions posées à une seule exception : la sécurité (figure 6), seul point sur lequel le cahier des charges en santé est beaucoup plus exigeant qu'en biologie ou en chimie.

Sur la base des résultats du sondage et de l'analyse, le groupe de travail a produit des recommandations décrites dans le chapitre suivant.

2. RECOMMANDATIONS GÉNÉRALES

2.1 Nécessité du développement de la grille

Le besoin d'une grille de production a été clairement exprimé par les personnes sondées. Plus de 40% des personnes sondées expriment un besoin important (entre 1 et 10 ans CPU) à très important (supérieur à 10 ans) de ressources de calcul sur cluster ou sur grille. Pour la biologie, l'émergence de nouvelles technologies expérimentales à haut-débit permettent de poser de nouvelles questions (biologie des systèmes, métagénomique, e-Cell…) nécessitant le recours à des moyens de calculs et de stockage massifs. La communauté bioinformatique (ReNaBi) souhaite anticiper ces nouvelles demandes de la communauté des biologistes.

La communauté utilise aujourd'hui plusieurs infrastructures de production avec à la clef une production scientifique significative :

- la grille Décrypthon,

- la grille EGEE et notamment l'organisation virtuelle Biomed qui permet d'accéder à plus de 20.000 processeurs dans le monde entier,

- des infrastructures de grille régionale (AuverGrid, GRIF),

- les ressources et services du Centre de Calculs de l'IN2P3,

- une infrastructure distribuée de ressources bioinformatique au niveau du réseau ReNaBi (13 plateformes avec chacune une centaine de processeurs en moyenne). La fédération de ces ressources dans une grille permettrait aussi une économie d'échelle au niveau matériel et en ressources humaines, tout en répondant aux nouvelles problématiques de la biologie à grande échelle.

La mise en place d'une infrastructure de grille nationale répond au besoin d'intégrer l'ensemble de ces ressources et de développer des services pour la communauté.

2.2 Aspects opérationnels

Les communautés en biologie et en santé ont des contraintes opérationnelles différentes :

- dans le domaine de la biologie, un réseau national de centres de ressources bioinformatiques s'est structuré depuis 2004. Ces centres de ressources fédérés aujourd'hui au sein du Réseau National de Bioinformatique (ReNaBi) sont des candidats naturels à devenir les nœuds de la grille de production nationale à condition de disposer d'un renfort en ressources humaines pour administrer les sites et accompagner les utilisateurs,

- la communauté de recherche médicale ne dispose pas de l'équivalent du Réseau National de Bioinformatique. Les laboratoires disposent de peu de ressources humaines pour les tâches d'administration de ressources informatiques. Il faut donc privilégier le rapprochement entre laboratoires de recherche clinique et laboratoires de recherche en informatique et imagerie médicale afin de pouvoir renforcer ces regroupements en ressources humaines pour administrer les sites et accompagner les utilisateurs.

2.3 Extension géographique

Comme indiqué précédemment, les nœuds potentiels d'une grille de production pour la biologie et la santé peuvent d'ores et déjà être trouvés parmi les nœuds actuels des grilles Décrypthon et EGEE au titre de services génériques. Pour la Biologie/Bioinformatique, les plates-formes bioinformatiques nationales (certifiées par les coordinations nationales RIO et GIS IBISA) sont fédérées depuis 2004 au sein du réseau national ReNaBi (http://www.renabi.fr). Ces 13 plateformes nationales constituent dès à présent les pôles de compétences et de ressources informatiques (données et calculs) en bioinformatique pour la communauté scientifique. De plus ReNaBi a conduit une initiative de grille depuis 2007 pour permettre la mutualisation des ressources de ces 13 plateformes nationales pour la communauté bioinformatique nationale et européenne. Cette initiative GRISBI (GRIlles Support pour la BIoinformatique) a été reconnue par le GIS IBiSA, par l'octroi d'une labellisation en 2008 en tant que plateforme nationale.

2.4 Les besoins de formation et d'information

Plus de 20% des personnes sondées disent tout ignorer des grilles. Si l'on tient compte en plus du fait que les personnes ayant répondu sont à priori les personnes les plus sensibilisées, il apparaît clairement un très gros besoin d'information et de formation.

Nous estimons approximativement à 1000 le nombre d'utilisateurs d'ici 4 ans, dont 200 sont déjà formés. Il nous faudrait donc former 800 utilisateurs, en recensant environ 10 formateurs dans ce domaine.

2.5 Les passerelles vers la recherche sur les grilles

Plusieurs collaborations sont aujourd'hui actives entre des équipes de recherche en informatique actives sur la grille Grid'5000 et des équipes impliquées dans le déploiement d'applications biomédicales sur grille de production, que ce soit sur Décrypthon, sur EGEE ou sur les grilles régionales. L'existence et la vitalité des passerelles vers la recherche viennent du fait que le domaine est l'un des plus exigeants pour ce qui concerne les fonctionnalités du système d'exploitation de la grille et qu'il est donc gourmand d'innovation.

2.6 La grille au sein de l'«écosystème»

2.7 La poursuite des groupes de travail

La poursuite du groupe de travail va dépendre de la suite qui sera donnée à ce travail de prospective.

2.8 Aspects internationaux

La communauté française des utilisateurs d'infrastructures de grille en biologie et en santé est l'une des plus avancées au monde. Les laboratoires français ont joué un rôle important dans les principaux projets de grille européens du domaine dans les précédents programmes-cadres et continuent de la faire dans le 7ème PCRD. Présents dans plusieurs projets de préparation des futures infrastructures européennes, notamment EGI, ELIXIR et LifeWatch, ils y défendent des solutions technologiques interopérables pour que la communauté française puisse pleinement s'appuyer sur les ressources de la grille nationale de production.

3. PLAN ACTION

Comme nous l'avons souligné précédemment, la communauté biomédicale est déjà très active sur les grilles de production (AuverGrid, Décrypthon, EGEE, GRIF). Cette activité va continuer et bénéficiera pleinement de la dynamique qu'apportera le déploiement de la grille de production nationale.

3.1 Actions à court terme

A court terme, la communauté va continuer de se structurer autour des grilles de production existantes, mais aussi poursuivre les collaborations avec les équipes de recherche en informatique exploitant Grid'5000. Elle poursuit aussi activement son action de lobbying dans les projets européens de définition des infrastructures de recherche.

Pour accompagner et répondre aux problématiques de la biologie à grande échelle, la plateforme GRISBI envisage dans les 2 années à venir l'interconnexion des 6 plateformes initiales en tant que grille avec un souci d'élargissement aux autres plateformes, et d'interopérabilité avec les grilles existantes. Cette première expérience permettra de tirer des conclusions sur une grille à plus large échelle pour la biologie.

3.2 Ressources humaines

La plupart des ressources humaines disponibles aujourd'hui pour l'administration des sites, le développement et le déploiement d'applications biomédicales sur les grilles sont issues de projets régionaux (AuverGrid) nationaux (ANR) et européens (EGEE, EMBRACE). Seuls quelques rares ingénieurs ont pu être recrutés sur des postes permanents et leur nombre est totalement insuffisant pour accompagner la migration de la communauté vers la grille de production. Le besoin prioritaire unanimement identifié pour généraliser l'adoption des grilles est le recrutement d'ingénieurs, de bioinformaticiens et de neuroinformaticiens ou d'informaticiens dans le domaine de la santé qui servent d'intermédiaire entre les utilisateurs finaux et les équipes qui conçoivent les middlewares et administrent les sites de production. En conclusion, l'adoption généralisée des grilles par la communauté de recherche en biologie et santé doit s'appuyer sur le recrutement d'une dizaine d'ingénieurs dont quelques-uns des ingénieurs expérimentés déjà présents dans la communauté sur postes permanents.

3.3 Moyens matériels et financiers

Pour la biologie, la plateforme GRISBI se propose dans un premier temps de fédérer les ressources matérielles disponibles du réseau ReNaBi. Dans ce sens, les demandes concerneraient en priorité le fonctionnement (frais de mission, ateliers, colloques, réunions…) et les besoins en formations pour les membres de GRISBI, et le transfert aux biologistes.

Pour avoir un ordre de grandeur des besoins matériels, il est intéressant de noter que l'organisation virtuelle Biomed consomme environ 5% des ressources de la grille EGEE depuis plusieurs années. Sur la base de ces chiffres, on peut estimer le besoin de la communauté française en ressources de calcul et de stockage en 2008 à 900 kSI2k et 1000 TB et en 2012 à 4500 kSI2k[1] et 3000 TB[2].

3.4 Gouvernance

La communauté de recherche ne souhaite pas développer une grille indépendante pour les sciences du vivant et de la santé mais s'inscrit résolument comme partenaire de la mise en œuvre d'une ou plusieurs grilles pluridisciplinaires dans la mesure où celles-ci offrent toutes les garanties en termes de stabilité et de pérennité.

L'adoption de la grille de production est très étroitement liée au fait qu'elle offre les services spécifiques dont la communauté a besoin, notamment en termes de sécurité, et qu'elle déploie des solutions technologiques interopérables avec celles mises en œuvre dans les infrastructures de recherche qui vont structurer la communauté au niveau européen.

[1] Ksi2k = 1000 SpecInt 2000. Les cœurs de processeur utilisés couramment ont une puissance variant entre 1,5 et 2,5 ksi2k suivant leur nature et leur fréquence d'horloge.

[2] TB = Tera Byte ou Téra Octets = 1000 Giga Octets

Monique Petitdidier, Franck Le Petit, Sophie Godin-Beekmann, Pierre Le Sidaner, Jean-Pierre Vilotte, Stratis Manoussis, Geneviève Moguilny, Karim Ramage, David Weissenbach

1. BILAN DE L'INTÉRÊT SCIENTIFIQUE

Plusieurs actions ont été entreprises au sein de la communauté Planète-Univers dans le cadre de la prospective de l'Institut des Grilles. Elles ont permis d'une part de présenter les Grilles à cette communauté et d'autre part de dresser un panorama de l'utilisation et des besoins sur les Grilles de production dans les communautés Sciences de la Planète et Sciences de l'Univers. Ces besoins s'inscrivent dans les spécificités de ces deux communautés.

1.1 Les communautés Sciences de la Planète et Sciences de l'Univers

Les deux communautés, Sciences de la Planète et Sciences de l'Univers, ont de nombreux points communs. Toutes deux étudient des phénomènes naturels sur lesquels l'Homme ne maîtrise aucun paramètre (météorologie, éruption volcanique, tremblement de terre, évolution des galaxies, …). Les deux principaux moyens d'investigation sont les observations et la simulation numérique.
Dans ces communautés, les observations sont précieuses étant donné leur caractère unique (cf. importance de la climatologie), pas seulement parce que des événements sont exceptionnels, (comme les éruptions volcaniques ou l'explosion de supernovae), mais aussi à cause du coût qu'elles représentent. En effet, les missions permettant l'acquisition des données se chiffrent en milliers jusqu'à des centaines de millions d'euros (mission satellites, grands télescopes, mission sur le terrain avec déploiement d'instruments sol et aéroportés). Les deux communautés ont besoin d'archiver les données acquises, d'y accéder régulièrement et de les exploiter avec des ressources de calcul adaptées au volume de données et aux algorithmes utilisés. Cette organisation autour des données a conduit au développement, au niveau international de standards d'échange et d'interopérabilité déployés au dessus des centres de données (Ex : Observatoire Virtuel en astronomie, données satellitales).

Les simulations numériques sont également au cœur de la recherche. Elles constituent des laboratoires in silico pour étudier les processus physiques. Les besoins en moyens de calcul pour les deux disciplines vont de l'ordinateur individuel au super-calculateur. Les communautés Sciences de la Planète et Sciences de l'Univers sont parmi les plus importants consommateurs de ressources informatiques sur les grands centres de calcul nationaux. Les mésocentres, équipement le plus proche des machines que l'on trouve sur les Grilles de production, rassemblent des ressources de calcul très utilisées par ces communautés pour les simulations et l'exploitation des données.

Dans les deux communautés, les grands projets se font la plupart du temps entre équipes géographiquement dispersées. Le partage des données est donc nécessaire. Ils mettent en relation des équipes de recherche publiques, parfois privées (en particulier en Science de la Planète) ainsi que des agences et des organismes (Météo France, CNES, ESA, NASA).

Un point clef des deux communautés est leur hétérogénéité et leur interdisciplinarité. Les équipes de recherche en Sciences de la Planète et en Sciences de l'Univers sont spécialisées sur des thématiques particulières : Atmosphères, Océan, Terre d'un côté, Cosmologie, Galaxies, Physique stellaire, Milieu interstellaire, Planétologie de l'autre. Les recherches scientifiques dans les deux communautés font appel à de nombreux domaines de la physique. Les projets rassemblent des équipes de ces communautés avec des expertises complémentaires mais également des équipes d'autres disciplines. Le vaste projet d'observation AMMA par exemple a rassemblé autour de Météo France des équipes travaillant sur l'océan, l'atmosphère, l'hydrologie, l'agriculture, la santé … De cette multi-disciplinarité, il résulte que les moyens et les outils numériques nécessaires aux équipes de recherche de ces deux disciplines pourront être très différents d'une équipe à l'autre, et d'une phase de leur projet à l'autre.

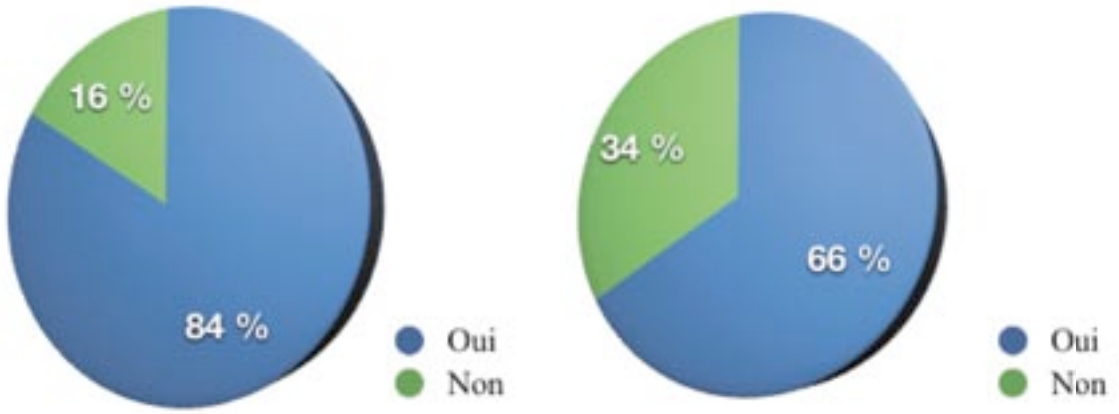
2. UTILISATION DES GRILLES DE CALCUL, BESOINS ET VERROUS

À l'exception de quelques équipes, les communautés Sciences de la Planète et Sciences de l'Univers utilisent peu les Grilles.

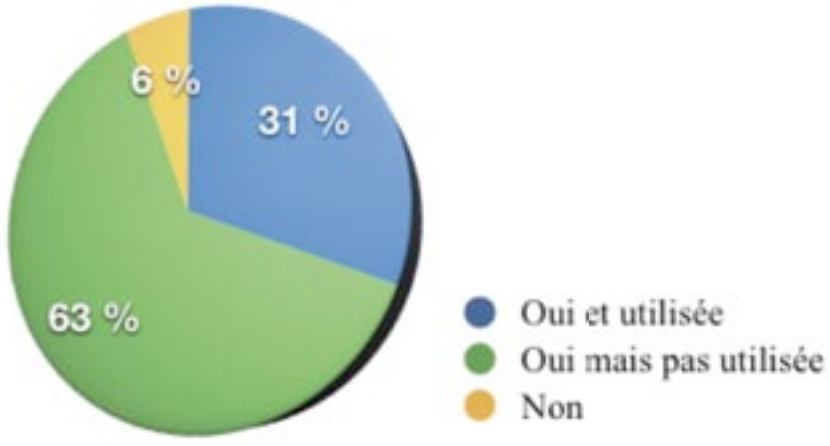
Quelques équipes de la communauté Science de la planète disposent déjà d'une forte expérience du calcul sur Grille comme l'IPSL et l'IPGP qui ont participé aux différentes phases d'EGEE. Le projet Européen DEGREE (Dissemination and exploitation of Grids in Earth Science) a permis la rédaction d'un livre blanc concernant la stratégie pour l'adoption de la Grille par une large partie de cette communauté : http://www.eu-degree.eu/DEGREE/internal-section/wp6/DEGREE-D6.1.2_v2.8.pdf/view.

Dans la communauté Astronomie et Astrophysique (A&A), les Grilles de production sont pour l'instant peu utilisées mais suscitent un intérêt pour certains problèmes, en particulier en simulation numérique. L'INSU a financé en 2008 un premier nœud EGEE dans la communauté Astronomie et Astrophysique. Plusieurs actions en 2008 ont été entreprises pour permettre aux astronomes et astrophysiciens de découvrir ce nouvel outil.

Dans le but d'établir un état des lieux sur les besoins en Grilles de production dans les deux communautés, un questionnaire a été diffusé à l'ensemble des membres des Sciences de la Planète et Sciences de l'Univers. Une fraction significative des personnes ayant répondu se disent intéressées par ce nouveau mode de calcul. Les équipes scientifiques y voient un moyen de subvenir au manque de ressources de calcul et de stockage pour des applications comme le traitement massif de données ou l'exploration d'espaces de paramètres pour lesquelles la Grille est particulièrement adaptée. Ils y voient un moyen de travailler plus rapidement, éventuellement sur des projets faisant intervenir des équipes géographiquement dispersées, et de s'attaquer à des problèmes plus complexes que ce qui peut être fait avec des ressources de calcul classiques.

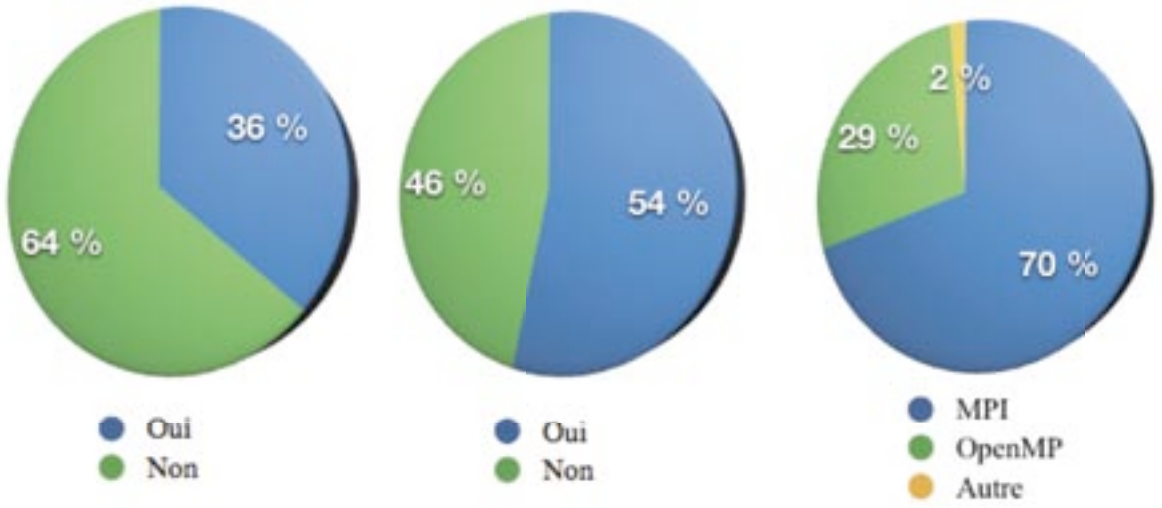


Réponse aux questions : A gauche : Travaillez-vous avec des applications dans lesquelles vous devez exécuter un grand nombre de fois un programme en faisant varier ses paramètres ? A droite : Avez-vous à utiliser ou produire de grandes masses de données ou appartenez-vous à des projets impliquant un partage de grandes masses de données avec des collègues distants ?



Réponse à la question : Verriez-vous un intérêt à l'utilisation d'une grille de production ?

Si, malgré cet intérêt la Grille est sous-utilisée dans les communautés des Sciences de la Planète et de l'Univers, c'est essentiellement dû à une méconnaissance de cette technologie. Les utilisateurs potentiels hésitent à s'investir dans la Grille tant qu'ils ne sont pas certains qu'elle puisse répondre à leurs besoins. L'utilisation quotidienne nécessite ensuite qu'ils puissent trouver rapidement réponses à leurs questions, ce qui suppose la mise en place d'un réseau d'experts de la Grille dans les communautés Sciences de la Planète et de l'Univers. Enfin, près de la moitié des personnes ayant répondu au questionnaire souhaitent utiliser la grille pour des applications parallélisées (OpenMP et MPI) ou utilisant des logiciels sous licence. Par conséquent, pour être adoptée, la Grille devra pouvoir intégrer les environnements communément utilisés dans nos communautés.



Réponse aux questions : 1 - Vos applications utilisent-elles un logiciel sous licence ? 2 - Vos applications sont-elles parallélisées ? 3- Quel type de parallélisme ?

Au niveau des communautés, le principal verrou est le lien entre la Grille et la structuration nationale et internationale des communautés. Celles-ci se sont organisées autour des données avec des centres de données hors de la Grille et des projets d'interopérabilité / accès aux donnés suivant des normes définies au niveau international. Si la Grille a le potentiel de permettre un traitement plus rapide de ces données, elle ne sera globalement acceptée que si elle est compatible avec l'organisation actuelle des Sciences de la Planète et Sciences de l'Univers.

3. CONCLUSIONS & RECOMMANDATIONS

3.1. Nécessité du développement de la grille

Face à l'exploitation systématique des données archivées qui ont été peu exploitées jusqu'à présent, à de nouveaux instruments d'observation produisant toujours plus de données à traiter et à des simulations pour étudier plus finement les processus physiques, l'accès quotidien à des ressources informatiques importantes est une nécessité pour les équipes de recherche des Sciences de la Planète et de l'Univers. La Grille, plus que tout autre équipement informatique, est particulièrement adaptée à un certain nombre d'applications comme ce qui concerne l'exploration des espaces de paramètres ou le traitement massif de données. Le développement de la Grille pour offrir aux chercheurs les moyens informatiques qui leur permettront de répondre aux nouveaux enjeux scientifiques est une nécessité. Un des aspects important est aussi son potentiel pour la e-collaboration (partage de données, de résultats et d'outils).

3.2. Aspects opérationnels

Le nombre d'utilisateurs des Grilles de production peut exploser dans les prochaines années. Cependant, elle ne sera acceptée que si son développement se fait en prenant en compte les spécificités des communautés des Sciences de la Planète et des Sciences de l'Univers :

- elle doit supporter les environnements de travail les plus courants des communautés en terme de logiciels, langages, interfaces avec des sites de calcul et de données externes…
- elle doit être compatible avec la structuration de nos communautés en terme de centres de données et d'accès aux données dont les normes sont définies internationalement.
- son utilisation quotidienne ne se fera qu'à condition qu'elle soit stable et fiable dans le temps.

3.3. Extension géographique

Si la Grille intéresse les communautés Sciences de la Planète et Sciences de l'Univers, à part pour quelques équipes, son utilisation est encore modeste. Cette faible utilisation provient d'une méconnaissance de la Grille et est imputable au manque de contacts entre les experts de la Grille et les communautés Sciences de la Planète et Sciences de l'Univers, une des raisons, pas la seule, est qu'il n'y a pas toujours de sites de grille dans des villes-pôles de PU.

L'expertise grille se trouve naturellement là où sont localisés les points d'accès à la Grille. Afin que les communautés acquièrent le savoir-faire et qu'ils puissent l'utiliser au quotidien, il est nécessaire de profiter du développement de la Grille pour mettre en place un réseau territorial de spécialistes de la Grille dans les deux communautés. Il serait souhaitable que les principaux OSUs (Observatoire des Sciences de l'Univers) qui existent en astronomie, géophysique et océanographie, disposent d'un nœud Grille avec du personnel formé.

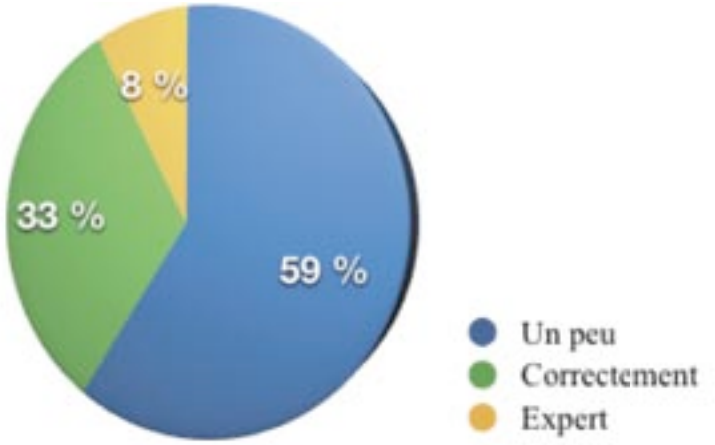
3.4. Besoins d'information et de formation

L'organisation de réunion d'informations, la participation à des événements de communication comme la ville Européenne des sciences, et surtout la présentation de résultats scientifiques obtenus grâce à la Grille sont autant d'incitations à l'attention des scientifiques pour qu'ils prennent le temps de s'intéresser à ce nouveau mode de calcul.

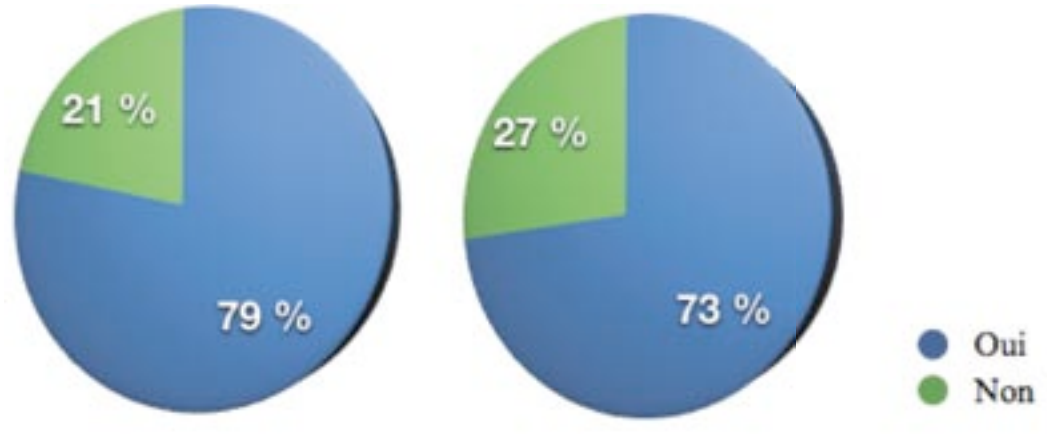
Un verrou important est l'acquisition du savoir-faire par des informaticiens et scientifiques. L'organisation d'une formation continue est indispensable. Les formations doivent être dirigées vers :

- des experts locaux qui aideront ensuite les utilisateurs à porter leurs applications sur la Grille de façon à avoir un mode de fonctionnement comme il en existe dans les centres de calcul.
- les scientifiques pour leur apprendre à utiliser ce nouveau moyen de calcul et en découvrir les potentialités.

Le nombre de formateurs et de formations nécessaires est difficile à estimer au-delà du court terme. En effet, si la Grille répond aux attentes des communautés Sciences de la Planète et Sciences de l'Univers, le nombre d'utilisateurs peut exploser très rapidement. Il faudra tenir compte de l'organisation régionale et des demandes des participants qui auront à la fois besoin de tutoriel de base pour les nouveaux venus et de formation plus spécialisée et/ou appliquée.



Pensez-vous êtres suffisamment informé sur les grilles de production et leur apport potentiel à votre recherche ?



Réponse aux questions : (gauche) Souhaiteriez-vous assister à un séminaire d'introduction aux grilles de calcul ? (droite) Etes-vous prêts à suivre une formation de quelques jours sur les grilles de production ?

3.5. La grille au sein de l'infosystème

Au sein des Sciences de la Planète et de l'Univers, les Grilles de production ne peuvent être qu'un moyen de calcul parmi d'autres. Les deux communautés continueront d'avoir besoin de super-calculateurs, de mésocentres, de moyens locaux et de Centres de Données hors de la Grille. La Grille a cependant toute sa place à prendre pour permettre aux scientifiques de traiter de nouveaux problèmes et de partager plus efficacement des données au sein de projets nationaux ou internationaux. Cela signifie qu'il est nécessaire de créer des interfaces pour passer de ces moyens de calcul à la Grille selon les besoins.

Les communautés Sciences de la Planète et Sciences de l'Univers se sont structurées autour de centres de données et ont développé des standards et des protocoles d'accès. Cette structuration et ces développements se sont fait hors de la Grille souvent au niveau international. Il est par conséquent nécessaire de réfléchir aux interfaces entre ces systèmes informatiques et les Grilles de production. Cette problématique a été / est abordée dans plusieurs projets européens (Open Geospatial Consortium, Global Earth Observation System of Systems, EuroVO-DCA, Open Grid Forum). Il est souhaitable que le développement d'une infrastructure Grille nationale prenne en compte ces spécificités et contribue à y apporter des solutions.

3.6. La poursuite des groupes de travail

Le groupe de travail «Sciences de la Planète et de l'Univers » mis en place dans le cadre de la prospective de l'Institut des Grilles a permis d'établir des collaborations entre les deux communautés. Afin de structurer les efforts dans le cadre de la gouvernance Grille pour ces communautés et de renforcer les collaborations naissantes, il semble judicieux de poursuivre l'existence du groupe de travail. Son rôle et sa structure restent à définir dans le cadre d'une NGI et devront être complémentaire des autres organisations (Actions spécifiques du CNRS,…) déjà en place sur cette thématique.

3.7. Aspects internationaux

Les projets sont de plus en plus internationaux. Par conséquent, la grille nationale doit être interopérable avec les grilles de nos partenaires, ce qui est le cas à l'heure actuelle avec EGEE. La structuration dans le cadre d'EGI qui sera mise en place devra tenir compte de ces collaborations internationales.

4. PLAN ACTION

4.1. Actions à court terme

Il semble nécessaire :

- de favoriser le déploiement de nœuds de Grille de façon à disposer d'un réseau de centres sur lesquels les utilisateurs pourront accéder à la Grille et trouver de l'aide pour son utilisation.

- d'organiser des réunions d'information sur la grille au niveau régional en particulier dans les villes où il n'existe pas déjà de centres de grille de production mais où d'importantes communautés des Sciences de l'Univers et/ou des Sciences de la Planète sont présentes (Bordeaux, Grenoble, Nancy, Rennes, Toulouse, …). Elles aideront les scientifiques à identifier les applications qui bénéficieraient d'un portage sur la Grille.

- d'organiser des tutoriaux de formation à l'attention des scientifiques et des ingénieurs en calcul scientifique des laboratoires, environ 2 à 3 par an.

- d'effectuer des portages concrets d'applications sur la Grille destinées à un groupe d'utilisateurs.

- d'aider, dans le cadre des actions déjà en court dans les deux communautés, à interfacer les centres de données et la Grille.

En 2009, il faudrait prévoir 5 à 6 réunions pour l'information et l'information.

4.2. Ressources humaines

Les ressources humaines sont indispensables pour l'expertise locale et aussi pour des développements concernant les interfaces entre les environnements couramment utilisés dans nos communautés et la grille. Nous avons cité précédemment l'interface avec les centres de données et les mésocentres.

Certains besoins en personnel peuvent être mutualisés avec d'autres disciplines en particulier la gestion des points de grille. D'autres ne peuvent pas l'être car ils sont associés à des besoins spécifiques à nos disciplines ou nécessitent une expertise proche des équipes scientifiques (comme c'est le cas en algorithmique lorsqu'il s'agit de développer des méthodes numériques novatrices qui tireront parti de la Grille).

Cinq à six recrutements dans les principaux centres régionaux PU sont indispensables dans les toutes prochaines années.

4.3. Moyens matériels et financiers

Il est difficile d'évaluer les moyens matériels et financiers à court terme. Cependant il faudra :

- Compléter l'équipement de points de grille déjà existant pour répondre à l'arrivée de nouvelles équipes et participer à leur fonctionnement ;

- Aider à s'équiper et mettre en route les nouveaux points de grille, en particulier dans des villes où rien n'existe ;

- Soutenir le fonctionnement des points de grille et la formation des personnels ;

- Accélérer l'acquisition des outils « Grille » en finançant des missions aux réunions prévues dans le cadre d'EGI de façon à ce que les différentes communautés se rencontrent et partagent leurs expériences et leurs outils de Grille.

4.4. Gouvernance

Nous proposons :

- de partir du groupe qui a travaillé sur la prospective et qui regroupait des personnes des 2 communautés ;

- de le structurer pour répondre aux besoins à moyen et long terme ;

- de lui associer de nouveaux membres pour être représentatif des besoins et des actions touchant aux Grilles déjà en cours dans le cadre d'autres projets.

Fairouz Malek, Tristan Beau, Jean-Pierre Meyer, Roman Poeschel et Cristinel Diaconu

Introduction

L'objectif de cette note est de faire la synthèse des expressions de besoins en moyen de calcul exprimés lors d'un sondage mis en place au mois de mai 2008. Le questionnaire portait, non seulement sur les besoins en stockage, en puissance de calcul et en connexions réseau, mais aussi sur le nombre d'utilisateurs, les collaborations et les problèmes rencontrés. Il s'agissait de dégager, pour chaque expérience, les grandes tendances aussi bien au niveau des besoins que des moyens utilisés et en particulier de préciser quelle fraction des ressources sera utilisée sur une grille de calcul dans les 3 à 5 ans à venir.

Le sondage a été adressé conjointement, aux utilisateurs du centre de calcul de Lyon (CC-IN2P3) en passant par le canal des responsables calcul et des responsables scientifiques des expériences, et aux directeurs d'unités de l'IN2P3 et du CEA/DSM/IRFU.

1. LES EXPÉRIENCES LHC

Il est sans doute utile de rappeler à ce stade que la stratégie des expériences LHC dans le traitement des données est clairement établie via le projet grille de calcul LCG («LHC computing Grid») qui concerne environs 360 chercheurs en France. Avec un taux de collision allant de 100 millions par seconde en 2008 à 1 milliard par seconde en 2010, le LHC va produire un flux de données jamais atteint jusque là. Grâce à l'électronique implantée dans les détecteurs, seule une petite fraction des collisions d'intérêt physique seront enregistrées (de 100 à 2000 Hz selon l'expérience). La production annuelle de données à traiter représentera de l'ordre de 10 à 15 Péta octets. Le projet LCG exploite trois grilles de calcul : OSG, EGEE, Nordu-Grid, et constitue aujourd'hui un réseau de plus de 200 centres de calcul situés sur 3 continents : Europe, Amérique, Asie. La grille LCG est d'ores et déjà opérationnelle (plus de 20 000 tâches traitées quotidiennement) et produit des données simulées et des données prises en tests avec des particules cosmiques permettant ainsi aux quelques 6000 physiciens de se préparer à la prise de données cet automne. Elle est hiérarchisée en quatre niveaux dont les trois premiers sont liés par un engagement formel de mise à disposition de ressources (MoU). Les centres ont les fonctions suivantes :

- le Tier-0 situé au CERN responsable de la pérennisation et de la distribution des données brutes provenant des quatre expériences LHC vers les 11 centres Tier-1.
- les Tier-1 (grands centres nationaux) assurent la reconstruction et la pérennisation de la fraction des données brutes qui leurs est confiée. Ils centralisent et distribuent les données réduites vers les 70 centres Tier-2 existant actuellement.
- Les Tier- assurent la production d'événements simulés et participent aux tâches d'analyses centralisées et stochastiques.
- Les Tier-3 sont des centres de ressources complémentaires aux centres Tier-2 assurant des tâches de simulation et d'analyse sur la base du volontariat, sans engagement vis-à-vis de LCG.

En France, l'IN2P3 et le CEA sont conjointement engagés dans le financement du Tier-1 au CC-IN2P3 qui sert les quatre expériences LHC et dont la contribution à l'effort Tier-1 est de l'ordre de 15%. À cela viennent s'ajouter une facilité d'analyse intégrée au CC-IN2P3, cinq centres Tier-2 et trois Tier-3 répartis sur la France dans les unités de l'IN2P3 et du CEA engagés dans les expériences LHC. La contribution Française à l'effort Tier-2 est de l'ordre de 10%.

2. RÉSULTATS DU SONDAGE

Parmi les 23 réponses collectées par le sondage :

- 9 proviennent de la communauté de physique des particules représentant environs 715 physiciens dont 360 appartenant à la communauté LHC (ATLAS, CMS, LHCb, ALICE).

- 5 proviennent de la communauté de physique nucléaire représentant environs 75 physiciens.
- 5 proviennent de la communauté des astro-particules représentant environs 240 physiciens.
- 3 proviennent d'utilisateurs du domaine biomédical et représente une dizaine de chercheur.
- 1 provient du domaine de l'archéologie et représente environs 5 chercheurs.

Dans nos présentations nous regrouperons les deux dernières catégories sous le nom générique « autre » dans la mesure où ils ne font pas partie de notre communauté mais sont néanmoins des utilisateurs du CC-IN2P3 et de la grille de calcul.

Globalement les réponses obtenues correspondent aux besoins de plus de mille chercheurs et à un volume en termes de calcul et de stockage de l'ordre de 85% des besoins de la communauté. On peut donc dire sans réserve que les données récoltées par ce sondage sont représentatives de l'ensemble de la communauté et que les chiffres mentionnés plus tard sont entachés d'une erreur ne pouvant pas excéder 10 à 15%.

Bien que les expériences LHC soient pionnières dans l'usage de la grille de calcul, environ 17% des physiciens travaillant sur des expériences hors-LHC utilisent d'ores et déjà la grille de calcul de manière routinière. Les autres disciplines, telle que l'astro-particule et la physique nucléaire n'ont pour le moment pas ou peu franchies le cap comme le montre la figure 1.

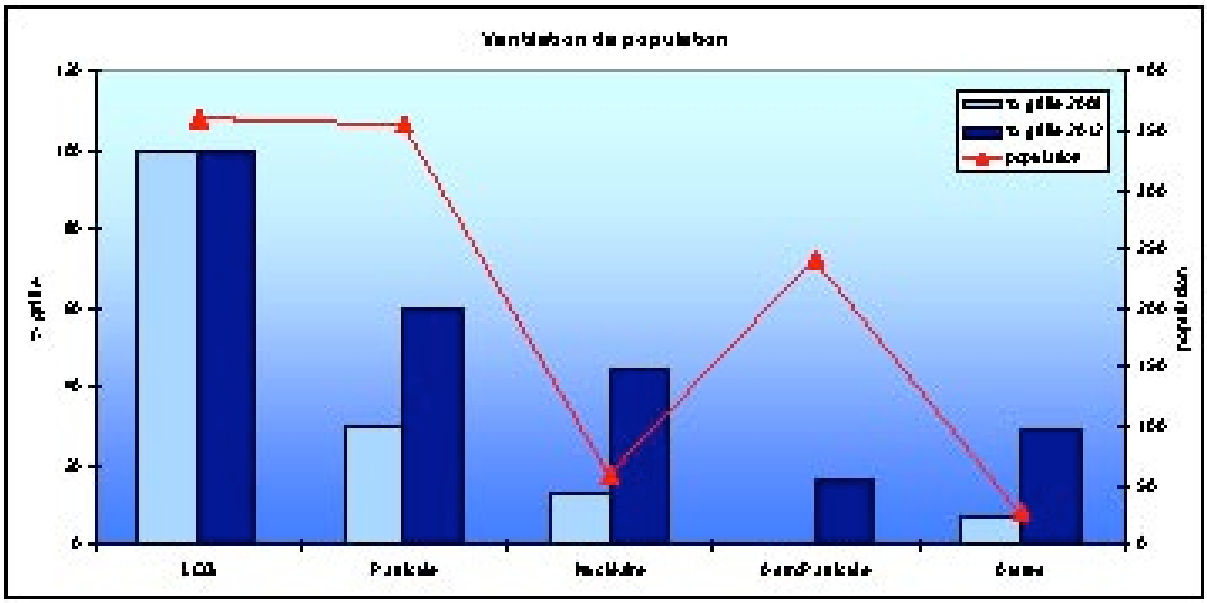


Figure 1 : Ventilation de population par thème de recherche. La ligne rouge représente la population de chercheur sondée par thème de recherche. Les barres bleues ciel représentent la fraction de population utilisatrice de la grille en 2008. Les barres bleues foncé représentent la perspective d'évolution pour 2012.

Globalement en 2008, 45% des chercheurs utilisent la grille de calcul régulièrement. En 2012 la fraction des chercheurs utilisant la grille devrait atteindre 65%. Parmi les utilisateurs potentiels actuels de la grille nous estimons à 30% la proportion des utilisateurs effectifs. Avec le démarrage du LHC et une migration des expériences vers la grille toutes disciplines confondues (45% vers 65%) nous pensons que le nombre d'utilisateurs effectifs de la grille va passer de 30% à 60%. Il faudra donc former à l'usage de la grille de l'ordre de 250 personnes sur les 4 années à venir. Les résistances principales à l'usage de la grille qui ont été exprimées sont les suivantes, par ordre décroissant d'importance:

- la complexité et la lourdeur d'emploi à comparer à un système déjà connu et satisfaisant pour certains groupes existants.
- Les instabilités et le manque de fiabilité dans l'accès aux données constatés lors de tests.
- Des problèmes d'interopérabilité principalement entre OSG et EGEE pour des expériences installées aux USA (D0 et STAR) et tournant pour le moment principalement sur OSG.
- L'inadaptation de la grille pour des applications nécessitant du calcul parallèle.
- Des problèmes liés à l'usage de licences payantes ou de calcul mono processeur de très grande durée (limitation à 36 ou 72 h selon les sites sur la grille).

L'inventaire des besoins en ressources de calcul et de stockage représenté sur la figure 2 montre que les deux tiers des ressources en 2008 sont utilisées par les expériences LHC. La fraction des ressources accédées en 2008 via la grille de calcul est de l'ordre de 73%.

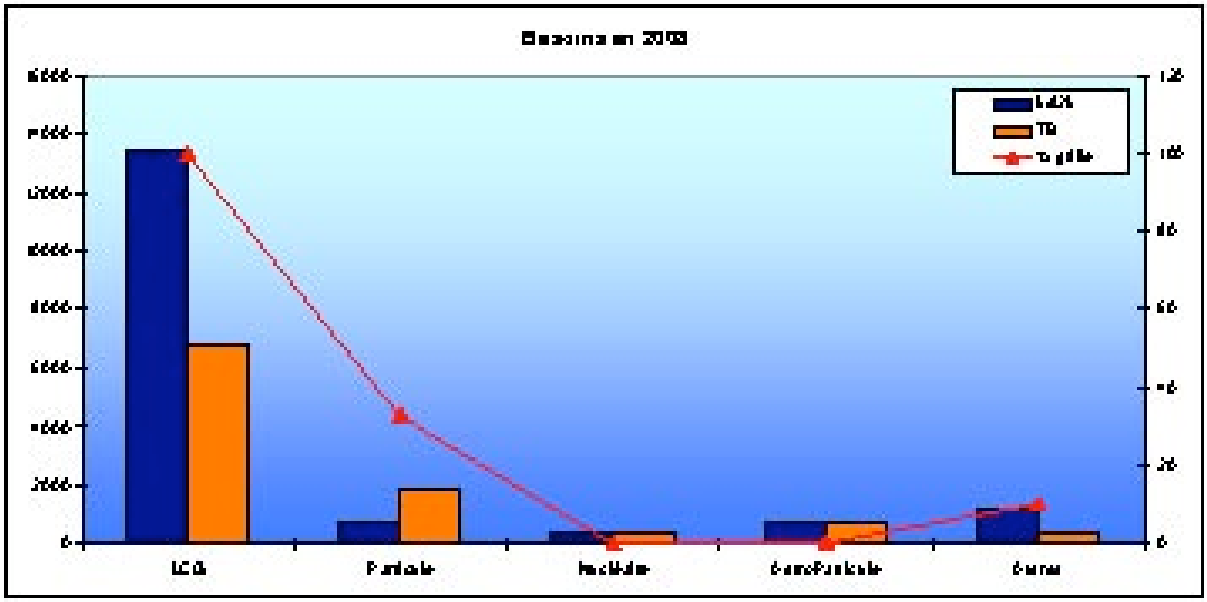


Figure 2 : ventilation des besoins en calcul (ksi2k[1] et barres bleues) et stockage (TB[2] barres oranges) en 2008 pour les chercheurs des différentes communautés sondées. La courbe rouge représente la fraction (en %) des ressources utilisées actuellement sur la grille.

En 2012, environ 80% des ressources seront dédiées aux expériences LHC (voir figure 3). La prospective montre que la fraction des ressources accédées via la grille de calcul, toutes disciplines confondues devrait avoisiner 85%. Il est à noter qu'en dehors des expériences LHC, la progression de l'utilisation de la grille proviendra principalement des domaines de la physique nucléaire et des astro-particules qui vont rattraper les chercheurs de la physique des particules, bien que les nouvelles expériences de physique des particules tel que l'ILC considèrent la grille comme le support principal.

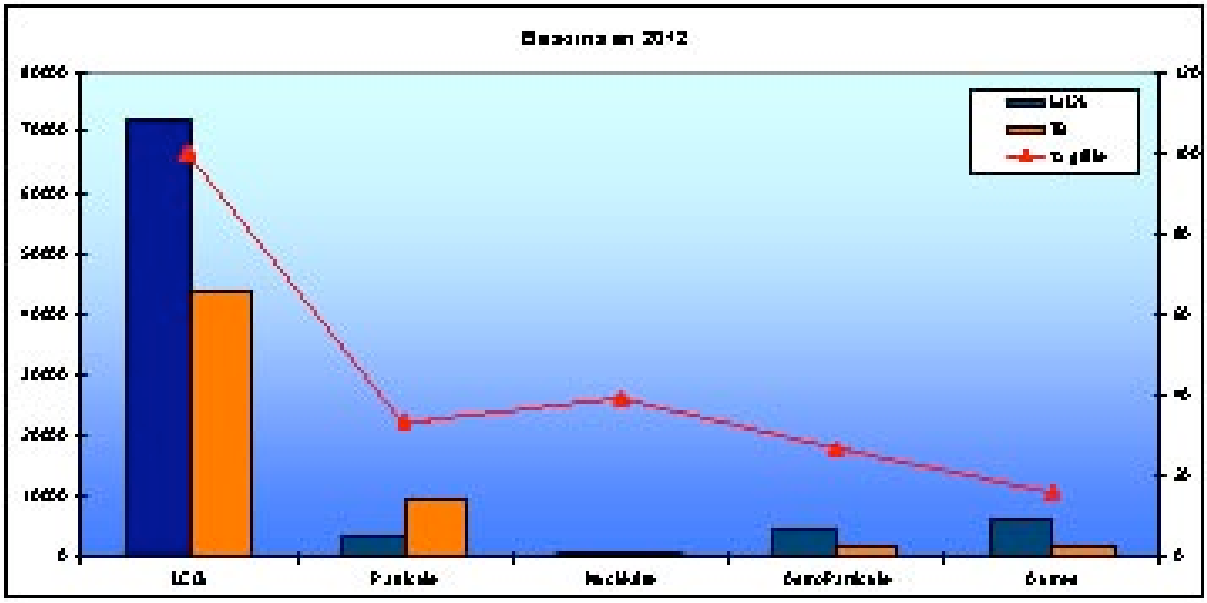


Figure 3 : ventilation des besoins en calcul attendue pour 2012. Les codes de couleurs employés sont identiques à la figure 2.

Pour finir, il est a noter que le réseau va devenir un outil très important pour les chercheurs avec de plus en plus d'application distribuées et des accès distant à des données et/ou des ressources de calcul. La tendance de croissance des besoins en réseau va varier d'un facteur 2 à 10 pour les laboratoires selon qu'ils hébergent ou non un nœud de grille.

3. RECOMMANDATIONS

La communauté de physique subatomique bien qu'historiquement pionnière sur les grilles de production devrait connaître dans les prochaines années, avec le démarrage du LHC et les études poussées de faisabilité de l'ILC, un accroissement continu des utilisateurs. D'autre part, le pouvoir attractif d'une grille de production arrivée à maturation et permettant de mutualiser facilement des ressources va dynamiser les utilisateurs des disciplines telles que l'astro particule et la physique nucléaire. Globalement en 2012 nous escomptons que 85% des ressources seront utilisées au travers de la grille de production. Dans le même laps de temps la communauté (environ 1000 chercheurs) passera de 45% d'utilisateurs actuellement à 65% soit un accroissement de 10% par an.

Nous recommandons donc que l'effort entrepris jusque là pour mettre sur pied une grille de production en France soit vigoureusement poursuivit avec une attention particulière pour les ressources dédiées au LHC afin de placer nos chercheurs dans une situation leur permettant de relever le défit de l'analyse des données du LHC dans un contexte de très forte compétition internationale. Les ressources LCG s'appuyant sur la grille EGEE, il est particulièrement important que la transition entre EGEE et EGI-NGI soit transparente aux utilisateurs. D'autre part certains problèmes résiduels d'instabilité, de fiabilité et d'interopérabilité doivent être solutionnés dans le cadre du consortium gLite prévu par EGI. Il faudra également trouver une solution permettant d'utiliser des licences de certains produits payants, tels que Mathematica par exemple, sur les grilles de production.

Philippe d'Anfray, Véronique Donzeau-Gouge, Cécile Germain-Renaud, Gaetan Hains, Michel Riveill, Alain Denise, Michel Beaudoin-Lafon, Xavier Pennec, Frédéric Desprez, Johan Montagnat, Olivier Richard, Eric Walter, Youssoufi Touré, Christian Saguez, André.De Lustrac, Stéphane Lantéri, Georges Cailletaud, Olivier Allix, Piotr Breitkopf

Introduction

Cette section couvre, sans prétention à l'exhaustivité, les domaines de recherche suivants : informatique, nanosciences et nanotechnologies, sciences des matériaux, mécanique des matériaux et des structures, automatique et traitement du signal, électromagnétisme et ondes, le calcul scientifique étant transversal. Il s'agit donc d'un ensemble très hétérogène.

L'informatique et les sciences de l'ingénieur (S2I) sont au cœur de la problématique, pas si nouvelle mais relativement récemment identifiée comme telle, de la simulation comme troisième pilier de la recherche scientifique, avec la théorie et l'expérience. Les communautés des sciences de l'ingénieur ont une position spécifique dans le développement, l'adoption et la diffusion de l'innovation dans le domaine du calcul informatique. Ici et par la suite, le terme « calcul » doit être entendu au sens large, incluant le traitement des données. La discipline informatique joue un rôle particulier, par ses composantes pour lesquelles les grilles ne sont pas une infrastructure de production, mais un sujet de recherche.

L'analyse détaillée montrera que ces communautés de recherche ont été exposées de façon très inégale à la problématique des grilles de production jusqu'à cette initiative. Un enjeu important de l'exercice était donc d'assurer une couverture suffisante des communautés, qui garantisse la pertinence des résultats. La composition du groupe de travail fournit de premiers éléments à cet égard, avec la participation des responsables de structures d'animations ou de pilotage de la recherche.

1. BILAN DE L'INTÉRÊT SCIENTIFIQUE

1.1. Les communautés et les grilles aujourd'hui

Quelques laboratoires d'informatique sont fortement impliqués dans deux grilles de production, EGEE et la grille support du programme DECRYPTHON[3]. C'est souvent, mais non toujours, dans un contexte pluridisciplinaire, en particulier en relation avec des applications dans le domaine de la biologie ou de l'imagerie médicale, comme le montre le chapitre portant sur les sciences de la vie. L'intervention de l'informatique se situe au niveau de travaux de recherche ou de leur application dans le domaine des systèmes distribués à grande échelle. Cette implication s'inscrit très souvent dans le cadre de programmes incitatifs (ACI, ANR, FP6 ou FP7). Quelques exemples marquants sont listés ci-dessous :

- AGIR[4] – Analyse Globalisée des données d'Imagerie Radiologique (programme ACI Masses de Données),
- DIET - Distributed Interactive Engineering Toolbox, un projet de recherche qui est aussi l'intergiciel de la grille DECRYPTHON,
- GWENDIA[5] - Grid Workflow Efficient Enactment for Data Intensive Applications (programme ANR Calcul Intensif),
- NEUROLOG[6] Software technologies for integration of process, data and knowledge in medical imaging (programme ANR TLOG),
- EDGES[7] enabling Desktop Grids for e-Science (FP7 IST Capacities programme),
- Grid Observatory[8] – Cluster d'EGEE-III et projet DIGITEO du même nom.

Le Groupe d'Utilisateurs Grilles et Calcul Intensif (Gus'G)[9] a conduit une activité d'animation, qui insère les grilles dans l'ensemble de l'écosystème allant du calcul intensif aux infrastructures de travail collaboratif. Une communauté informatique significative et active s'est donc créée.

Du côté des sciences de l'ingénieur, les grilles sont encore largement méconnues, mais un élément nouveau est intervenu dans la dernière période. Les actions d'animation et de diffusion d'EGEE à l'échelle internationale,

et aussi en France sur le campus d'Orsay, ont fait émerger une demande prioritaire de certaines communautés pour la disponibilité de logiciels de calcul scientifique, et spécifiquement Matlab. Le développement d'une solution complète par MathWorks pour EGEE[10] ouvre des perspectives absolument nouvelles.

- Elles offrent dès maintenant l'exploitation de la grille EGEE aux utilisateurs de Matlab.
- Elles ont permis d'acquérir dans GRIF (Grille de Recherche d'Ile de France, intégrée dans EGEE), une expertise technologique exploitable pour des développements analogues dans le domaine du logiciel libre.
- Elles ont crée un modèle nouveau de gestion de licences sur grille, qui peut servir de base pour le déploiement d'autres logiciels commerciaux.

Enfin, l'action de prospective a constitué par elle-même un outil de diffusion d'information, suscitant un intérêt et une attente qu'il importera de relayer.

1.2. L'enquête

Evaluer l'adéquation des grilles de production aux besoins des communautés concernées n'est pas immédiat. Contrairement par exemple au cas de la physique des particules, l'exploitation des infrastructures informatiques hors supercalculateurs n'y fait pas l'objet d'une planification, mais d'une prise de décision à l'échelle de laboratoires, de réseaux de recherche, voire de chercheurs individuels.

1.2.1. Le questionnaire

Le questionnaire et les résultats sont disponibles sur http://www.idgrilles.fr:spip.php?article44. L'objectif de l'enquête était de toucher les communautés au-delà des spécialistes des infrastructures informatiques, pour obtenir un sondage sur les besoins en général plus que sur les pratiques effectives à l'instant présent. L'enquête a donc privilégié des questions qualitatives et orientées vers les applications, et sous forme fermée. Une identification minimale (nom, email) était obligatoire. Les biais introduits par un exercice de sondage volontaires sont évidents ; toute l'analyse qui suit doit donc être interprétée en en tenant compte.

1.2.2. Synthèse des résultats

L'enquête a fourni 89 réponses, dont 24 émanent de responsables (d'équipes, de laboratoires, de structures d'animation). Quelques réponses proviennent de l'industrie. Les deux secteurs, informatique d'une part et sciences de l'ingénieur d'autre part sont représentés, mais les sciences de l'ingénieur ont répondu de façon plus importante, et souvent plus détaillée.

1.2.3. L'information

Assez naturellement, la communauté informatique se considère comme assez bien ou très bien informée, alors que le déficit d'information est criant dans les sciences de l'ingénieur (fig. 1). Il est cependant intéressant de remarquer qu'environ la moitié des réponses ont trouvé informatif le document d'introduction aux grilles de production rédigé par le groupe de travail. Ce document très synthétique a été élaboré à la suite de la première réunion du groupe de travail, qui a mis en évidence les points suivants, confirmés par la suite (enquête et échanges).

- La méconnaissance pure et simple de l'existence des infrastructures grilles de production.
- L'assimilation fréquente entre, soit grilles de production et grilles de recherche, soit grilles de production et environnements de calcul parallèle.

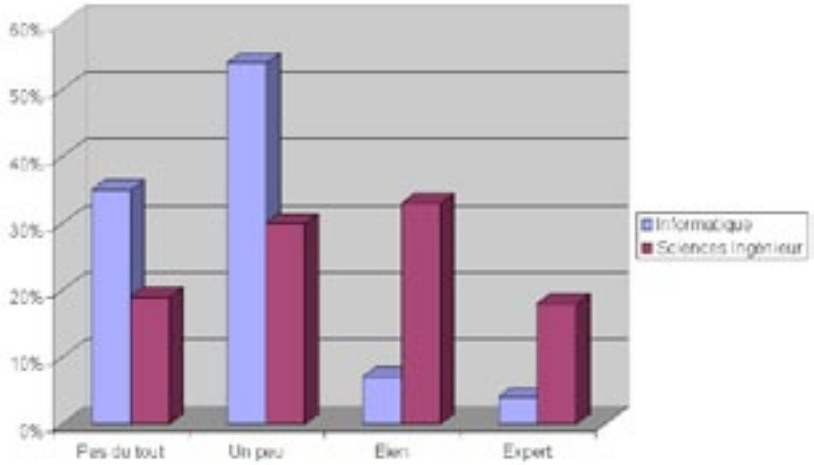


Figure 1 : Réponses à la question Pensez-vous être suffisamment informé sur les grilles de production et leur apport potentiel à votre recherche ?

1.2.4. Les applications

Dans les deux communautés, les applications multi paramètres sont dominantes. Le partage de grandes masses de données est marginal dans les sciences de l'ingénieur, un peu mieux représenté en informatique (fig. 2). Ce dernier point est un exemple typique du biais d'enquête : la communauté informatique nombreuse et active dans le domaine ne s'est pas sentie concernée.

L'enquête a fourni 35 « applications phares pour lesquelles l'accès à une grille de production procurerait un avantage immédiat et décisif ». Quelques exemples dans le domaine multi paramètres sont les suivants :

- Diverses applications de bioinformatique sont naturellement présentes.
- Plusieurs applications relèvent du domaine de l'optimisation et de la validation de méta heuristiques.
- Méthodes Monte-Carlo en électromagnétisme.

Une autre partie de ces applications est de type parallèle fortement couplé.

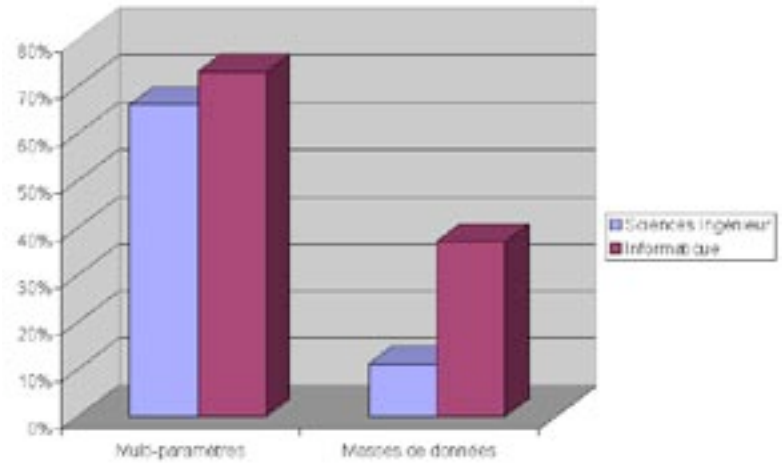


Figure 2 : Réponses (Oui/Non) aux questions :

Travaillez-vous avec des applications dans lesquelles vous devez exécuter un grand nombre de fois un programme en faisant varier ses paramètres d'entrée ?

Etes-vous partenaires de projets qui impliquent de partager de grandes masses de données avec des collègues distants pour un travail collaboratif ?

1.2.5. Les modes de production

Les réponses portant sur la limitation par les ressources disponibles ne sont pas nécessairement très significatives, l'offre n'étant pas toujours clairement connue. Environ la moitié des réponses font état d'une limitation, très majoritairement en calcul.

La communauté informatique a investi à la fois les modes de production grappes et grilles. On verra cependant dans l'analyse détaillée que la petite fraction de réponses qui souhaite exploiter ces modes de production, mais ne le fait pas encore, recouvre en fait une grande quantité d'utilisateurs potentiels. Pour les sciences de l'ingénieur, l'écart entre les modes de production souhaités et effectivement utilisés est nettement plus important.

1.2.6. L'intérêt pour les grilles de production

Cette question fournit un des résultats les plus nets de l'enquête : informatique (94%) et sciences de l'ingénieur (70%) voient un intérêt à l'utilisation d'une grille de production (fig. 3), ces chiffres étant consistants avec la réponse à la question suivante portant sur l'utilité. En revanche, l'exploitation actuelle de ces infrastructures est faible, ce qui est confirmé par une majorité de réponses percevant l'utilité comme certaine, mais non immédiate. Les réponses libres font apparaître clairement la difficulté à se projeter dans l'utilisation de ces infrastructures, liée à la fois au manque d'information, et aux spécificités du domaine S2I, qui seront discutées plus loin.

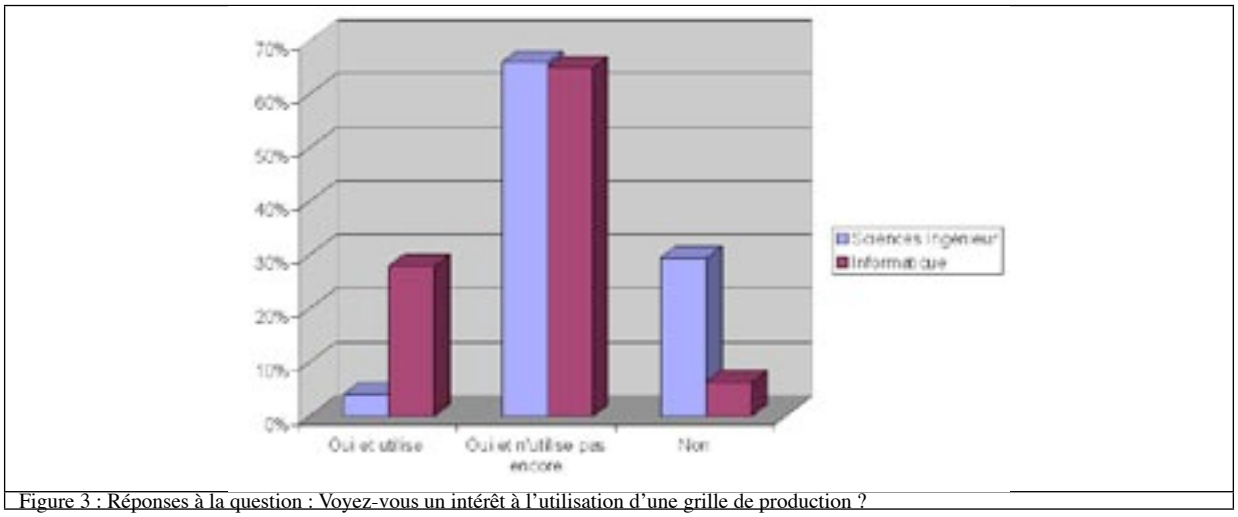


Figure 3 : Réponses à la question : Voyez-vous un intérêt à l'utilisation d'une grille de production ?

1.3. Les besoins par domaines

1.3.1. Informatique

Cette partie ne traite de la recherche informatique qu'en tant qu'utilisateur « banal » des grilles, qui l'utilise en tant qu'infrastructure. La section 3.3 traitera des interactions entre grilles de production et recherche sur les grilles.

Le besoin en puissance de calcul apparaît dans de très nombreux secteurs, avec souvent une forte composante pluridisciplinaire ou applicative qui motive le besoin de ressources élargies. Deux exemples parmi bien d'autres possibles, sont la bioinformatique, et l'apprentissage. Le niveau général d'exploitation est cependant marginal par rapport aux besoins réels.

1.3.2. Automatique et traitement du signal

Ces thématiques sont caractérisées par un usage massif de logiciels commerciaux de calcul scientifique. Si le verrou de la complexité d'accès à la ressource peut être levé, il y a au moins deux types de problèmes qui peuvent bénéficier d'une mise en œuvre sur grille de production. Le premier correspond aux études paramétriques : par exemple, études de performances statistiques de systèmes de communication ou pour l'optimisation de performances avec des méthodes de recherche aléatoire. Le second correspond aux problèmes de très grande taille qu'on rencontre, par exemple, en traitement d'images multi spectrales ou de séquences d'images et pour lesquels les ressources d'une station de travail se révèlent insuffisantes.

1.3.3. Calcul des structures

Une catégorie d'application pouvant bénéficier immédiatement des grilles de production est la simulation multiphysique (thermique - fluides - matériaux - tenue des structures - acoustique). Les industries impliquées appartiennent principalement aux secteurs aérospatial, automobile, énergétique, de transformation des matières premières, avec une forte présence de grands groupes.

Les actions de recherche (par exemple OMD, Optimisation Multi Disciplinaire[11]) sur des méthodes de conception collaborative et d'optimisation multidisciplinaire tentent d'aborder ce type de problèmes, mais vont se heurter inévitablement à la disponibilité de ressources informatiques suffisamment puissantes, ne serait-ce que pour l'établissement des plans d'expérience numériques pour des quantités de paramètres raisonnables. L'évolution d'OMD vers et OMD[11], Optimisation Multi Disciplinaire Distribuée, atteste du besoin de ressources élargies.

Dans les domaines Mécanique des matériaux et des structures, Microélectronique et Nanotechnologies, et Elecromagnétisme et Ondes, les besoins exprimés concernent surtout le parallélisme fortement couplé. Pour Microélectronique et Nanotechnologies, la demande porte aussi sur des versions parallélisées des environnements de calcul numérique commerciaux.

3 www.decrypthon.fr/

4 www.aci-agir.org

5 gwendia.polytech.unice.fr

6 neurolog.polytech.unice.fr

7 www.edges-grid.eu

8 www.grid-observatory.org

9 www-gusg.aristote.asso.fr

10 www.mathworks.com/company/pressroom/articles/article31250.html

11 http://omd.lri.fr/

2. RECOMMANDATIONS GÉNÉRALES

2.1. Modèles d'exploitation

Du point de vue opérationnel, pour les sciences de l'ingénieur, le support des logiciels commerciaux, et au premier titre Matlab (incluant l'ensemble des toolbox) est un prérequis. La grille EGEE est un support adapté pour ce schéma nouveau, qui donnerait un avantage compétitif très important aux équipes concernées. La grille est alors utilisée comme infrastructure de travail collaboratif (mutualisation de logiciel), l'horizon d'allocation de ressources étant le cluster. Le support à la diffusion et à l'exploitation de cet outil est une priorité.

Plus généralement, un point commun à la communauté S2I est un rapport spécifique entre développement et exploitation. Le cycle de vie des développements logiciels est souvent nettement plus court que dans d'autres communautés, la valeur ajoutée de la recherche étant dans les concepts nouveaux introduits dans le logiciel. Un consensus se dégage de l'exercice : la difficulté à investir dans l'apprentissage d'une technologie de production dont les avantages sont visibles, mais le retour sur investissement est perçu comme demandant un délai significatif, et la transmission de l'expertise aléatoire, en particulier avec le portage de logiciels existants développés dans des contextes très expérimentaux et pas toujours robustes. Les difficultés de la courbe d'apprentissage sont probablement surestimées, mais les écueils sont loin d'être uniquement techniques. Les décideurs et des utilisateurs ont avant tout besoin

- de lisibilité, pour identifier les échelles de temps et les coûts réels de mise en œuvre ;
- d'incitations à investir dans une technologie nouvelle.

De ce point de vue, un obstacle important à l'adoption des grilles de production est le manque de visibilité de celles-ci dans la programmation de l'ANR. Pour 2009, les programmes du domaine STIC, et en particulier COSINUS (Conception et Simulation) et ARPEGE, qui seraient le plus concernés par les grilles, n'en font aucune mention.

Une dernière conséquence est que les utilisateurs de cette communauté, qui, rappelons-le, jouent un rôle essentiel dans la diffusion de nouvelles technologies, ne sont pas toujours de gros consommateurs de ressources : ils expérimentent et créent des méthodes sans les exploiter au long cours. L'impact de leurs demandes sur les gestionnaires et décideurs des grilles de production risque donc d'être faible par rapport à celles des utilisateurs plus visibles. Le risque est alors de voir les utilisateurs potentiels se tourner vers des environnements moins efficaces, mais plus réactifs. Le rôle spécifique de la communauté S2I doit être mieux pris en compte. Dans le même esprit, l'extension des grilles de production vers un rôle de support du travail collaboratif, les réflexions prospectives et les expérimentations associées, doivent être intégrées à la réflexion stratégique sur le rôle des grilles de production.

2.2. Pour une formation et un support adaptés

L'exercice permet de constater une extrême disparité dans l'information et les pratiques, et ceci à toutes les échelles : dans un même laboratoire, peuvent coexister sans interaction experts et demandeurs de ressources.

La question cruciale est celle du suivi, de la formation et du support. L'existant d'EGEE-III dans le domaine de la formation est très riche. Le support, très internationalisé, ne répond pas à une situation où la formation et le support doivent avoir pour objectif de favoriser l'accès à la grille de petites équipes à la fois inexpérimentées et demandeuses de réactivité. Jusqu'à maintenant, les chercheurs en informatique impliqués dans EGEE ont pu jouer un rôle d'intermédiaire et d'orientation, mais ne sauraient suffire à la tâche telle qu'elle est révélée par l'exercice. L'expérience montre que le support et la formation sont assurés au mieux par des ingénieurs ou des chercheurs actifs dans des projets exploitant ou opérant la grille de production.

Sur le long terme, la meilleure garantie de l'exploitation d'une technologie est son introduction dans les cursus de formation. Si les aspects fondamentaux qui sous-tendent les grilles de production sont largement représentés dans les cursus d'informatique, l'intégration des grilles de production dans les formations dédiées à l'informatique hautes performances est rarissime. Cela aurait été prématuré jusqu'à une période récente, mais devient d'actualité dans les cursus d'informatique appliquée, et probablement aussi dans les formations aux outils informatiques des sciences de l'ingénieur.

2.3. Quelles interactions entre grilles de production et recherche sur les grilles ?

La discipline informatique joue un rôle particulier, par ses composantes pour lesquelles les grilles ne sont pas une infrastructure de production, mais un sujet de recherche. Des points de contact avancés de la recherche et de la production existent, et ont déjà abouti à de vraies réussites. Au-delà de ces actions spécifiques, le groupe de travail, en collaboration avec le GDR ASR (Architecture Systèmes Réseaux) a exploré les thématiques de collaboration plus générales, qui sont présentées sommairement ici et sont détaillées dans le chapitre 5 du document d'analyse du groupe de travail[12].

Comme pour les grandes disciplines scientifiques étudiant des systèmes complexes, l'observation de systèmes réels et la capacité à expérimenter sont des éléments fondamentaux de la méthodologie scientifique en informatique. La complexité des systèmes informatiques de demain, qui constitueront l'Internet du Futur, s'approchera de plus en plus de celle des systèmes biologiques. Cependant les chercheurs sont confrontés à la difficulté de réaliser des expérimentations.

Si aujourd'hui, plus personne ne critique la nécessité d'avoir accès à une grille de recherche, qui ne soit pas une grille de production, et dont tout peut être contrôlé et configuré selon les besoins du chercheur, les questions clé que sont l'acquisition de données, le processus de validation et les méthodes de transfert des résultats de recherche, les procédures de migration du logiciel, restent ouvertes. Des synergies ne peuvent être mises en place que si les deux communautés ont une bonne compréhension de leurs contraintes respectives.

Au-delà de ces généralités, des thématiques concrètes de collaboration, qui associent toutes une composante d'informatique fondamentale à une classe de problème opérationnels, ont été identifiées : grilles autonomiques ; extensibilité des ressources logicielles ; développement d'expérimentateurs croisés EGEE-gLite/Grid'5000 ; passage à l'échelle. Elles sont détaillées dans le document précité. La définition de ces thématiques constitue un premier pas pour répondre au souhait d'interactions scientifiques formulé dans le chapitre grilles régionales.

2.4. La grille au sein de l'écosystème

Nous n'abordons pas ici les relations entre grilles et supercalculateurs, traitées dans un autre chapitre. Un domaine émergent est celui des relations entre grilles d'entreprises, Cloud Computing et grilles de production. Dans ce domaine naturellement partenarial, l'ANR pourrait jouer un rôle essentiel, par l'ouverture de ses appels du domaine STIC aux projets visant les grilles d'entreprise. Cela nécessite un format adapté de financement mettant en partenariat des industriels fournisseurs ou utilisateurs de services organisés en grilles, des laboratoires publics en informatique pure et informatique de gestion, et des services ou ressources publics de grilles expérimentales.

2.5. Le groupe de travail

Cet exercice de prospective a suscité dans la communauté S2I un intérêt réellement important et nouveau : une « fenêtre » est ouverte, qu'il faut exploiter. Le groupe de travail y a joué un rôle essentiel. La forme actuelle n'a pas vocation à perdurer, mais le groupe pourrait participer au suivi et à l'impulsion des actions issues de l'exercice.

3. PLAN D'ACTION

3.1. Actions à court terme

- Support financier mutualisé pour la participation aux actions de formation d'EGEE, et à l'école d'été internationale sur les grilles (l'édition 2009 est organisée en France)
- Incitation et soutien à des actions de diffusion (présentation, démonstrations) de la grille EGEE dans les GDR.

3.2. Ressources Humaines

Ingénieurs de recherche

- Expérimenter, puis créer une référence, pour l'action Matlab-EGEE : 1 ingénieur de recherche. C'est la demande prioritaire.
- Pérenniser les projets existants (cf section 2) : 3 ingénieurs de recherche.

Chercheurs ou enseignants-chercheurs

- En informatique, les thématiques de recherche liées aux interactions entre grilles de recherche et grilles de production sont portées par des chercheurs expérimentés, mais ont besoin d'être renforcées par de jeunes permanents. Un poste par an (2009-2013) fléché sur ces thématiques.

Ces personnels contribueraient évidemment à l'action d'information et formation.

3.3. Moyens matériels et financiers

La présence de la thématique grilles de production dans des appels d'offres des programmes nationaux est un vecteur privilégié d'incitation. À moyen terme, une action équivalente à celle du programme ANR COSINUS serait la bonne échelle pour offrir à la communauté l'opportunité d'exploiter les infrastructures de grilles de production. La composante partenariale dans ce cas serait à rechercher du côté des utilisateurs d'une part, et de l'industrie du logiciel scientifique d'autre part. À plus court terme, la présence des grilles de production comme infrastructure cible dans les programmes est essentielle. Une action de même nature, à une échelle plus modeste, devrait cibler la communauté informatique sous le double aspect recherche et exploitation.

3.4. Gouvernance

Le caractère pluridisciplinaire de la gouvernance des grilles de production est unanimement considéré comme critique pour l'implication de la communauté S2I.

[12] http://indico.lal.in2p3.fr/getFile.py/access?contribId=20&resId=1&materialId=paper&confId=517

Carlo Adamo, Xavier Assfeld, Dorothée Berthomieu, Nicolas Ferré, Henry Chermette, Claude Pouchan, Michel Caffarel

Introduction

La Chimie s'attache de plus en plus à comprendre les principes essentiels mis en jeu dans les différents phénomènes chimiques complexes étudiés à l'aide des outils de la chimie théorique, comme en témoigne le rapport de conjoncture du CNRS de 2004.

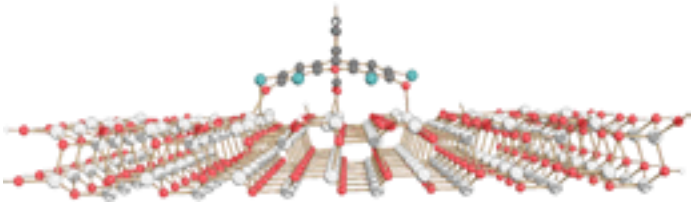


Figure 1 : Exemple de calcul en Chimie Quantique : modèle périodique à 2 dimensions pour la simulation d'une cellule photovoltaïque à colorants type Eosyne/ZnO. Le calcul de cette structure a requis 6 mois sur 8 processeurs bicœurs Opteron avec le code CRYSTAL (source équipe MSC, LECIME, UMR 7575).

Les récents développements informatiques (logiciels et matériels) ont permis de complexifier les modèles théoriques, ce qui a participé à la promotion de l'approche théorique auprès de la communauté des chimistes. Cette branche de la Chimie s'intéresse aux propriétés physico-chimiques d'un système, grâce à un modèle théorique (et chimique). De plus, elle permet de prédire le comportement d'un système, l'approche théorique pouvant même se substituer à l'expérimentation.

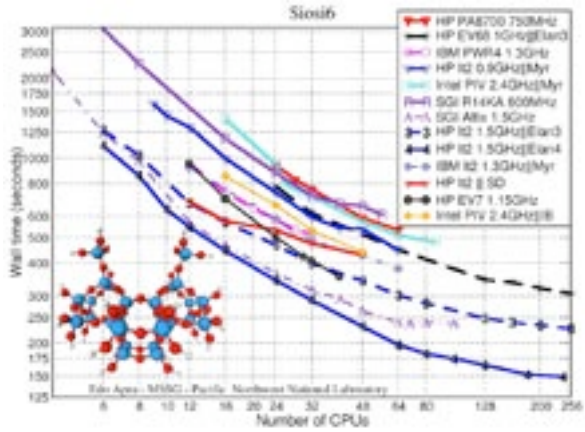


Figure. 2 :Performances du code NWChem (source : http://www.emsl.pnl.gov/docs/nwchem/nwchem.htm)

Ainsi, la Chimie Théorique (CT) est une discipline regroupant théorie, modélisation et simulation en Chimie. La CT agit à deux niveaux distincts. Le premier est celui de l'explication ou compréhension, dans laquelle on utilise des lois ou théories établies afin d'étudier à différentes échelles les propriétés de systèmes chimiques ainsi que leur évolution. Le deuxième est celui de la prédiction, dans lequel des phénomènes complexes, faisant intervenir un grand nombre de paramètres, sont modélisés ou simulés grâce à un modèle inspiré de leurs caractéristiques chimiques.

Dans ce cadre, les principaux enjeux concernent l'étude des propriétés physico-chimiques et de la réactivité de la matière dans différents états d'agrégation (molécules, solution, solide, surface…) ou phases et à des échelles de temps variables (de la fs à la ns). Les outils théoriques dont disposent les chimistes peuvent être regroupés en deux familles : les méthodes quantiques (Hartree-Fock (HF), post-HF, DFT) et les méthodes classiques (mécanique moléculaire, mésoscale, coarse-grained). De plus, leur mise en œuvre repose sur différentes techniques algorithmiques plus ou moins sophistiquées (minimisation, trajectoires, Monte Carlo, …) Toutes ces méthodes, qu'elles soient quantiques ou classiques, sont très «gourmandes» en puissance de processeur, de mémoire et en temps de calculs, ce qui amène souvent le chercheur à faire un choix entre la taille du système étudié, la précision de la propriété recherchée, la puissance de calcul à sa disposition et le temps à investir (ou disponible) dans le calcul.

La plupart des codes utilisés dans la communauté des chimistes sont commerciaux, avec des exceptions notables (par exemple CPMD, NWChem, GULP). En général, ces codes sont bien parallélisés, ayant ainsi de bonnes performances en fonction du nombre de processeurs (voir par exemple figure 2).

1. Utilisation des ressources de calcul en Chimie

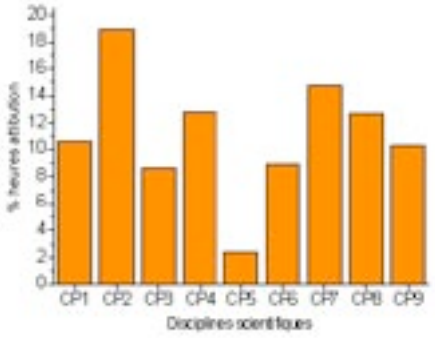


Figure 3 : Attribution des heures de calcul sur la machine Zahir de l'IDRIS en 2007 (source comité des utilisateurs de l'IDRIS).

Les calculs numériques sont donc au cœur de la recherche en Chimie, les différents types de problèmes étudiés et de méthodes employées (et donc de codes disponibles) étant très variés. Il n'est donc pas étonnant de constater que les besoins en puissance de calcul vont de la station de travail individuelle aux supercalculateurs. Plus précisément, la communauté Chimie est l'une des plus importantes consommatrices de ressources informatiques sur les grands centres de calcul nationaux (IDRIS et CINES). Ceci apparaît clairement sur la figure 3, dans laquelle sont représentées les heures attribuées en 2007 aux différents projets classés par discipline sur l'ordinateur Zahir de l'IDRIS. Cet exemple met en évidence que la Chimie, majoritaire dans le Comité de programme n° 8 (Chimie quantique et Modélisation moléculaire) et présente également dans le Comité de programme n° 9 (Physique, chimie et propriétés des matériaux), est le troisième utilisateur de ressources informatiques en France.

2. Intérêt des grilles en Chimie

Comme dans d'autres disciplines scientifiques, l'intérêt des grilles en Chimie réside à la fois dans une utilisation rationnelle des ressources informatiques à l'échelle nationale (voire européenne) et dans l'accès (ponctuel ou non) à des ressources allant bien au-delà de celles classiquement disponibles dans un laboratoire de recherche. Il faut aussi souligner que dans le contexte actuel où les superordinateurs sont de plus en plus spécialisés dans des applications ciblées (à cause de leur architecture), la grille représente un moyen « démocratique » d'accès à des ressources de calcul pour une communauté ne pouvant pas toujours supporter localement l'investissement nécessaire au calcul haute-performances.

Plus précisément, les chimistes sont intéressés par :

1) Le calcul intensif parallèle, dans lequel des grilles de calcul sont utilisées afin de combiner les capacités de plusieurs machines peu efficaces individuellement, permettant ainsi de résoudre un problème complexe généralement représenté par un système chimique de grosse taille (en termes de nombre d'électrons ou d'atomes). Ici, le parallélisme du code utilisé représente un aspect important, avec des codes caractérisés par une croissance linéaire même à nombre élevé de processeurs (par exemple NWChem, ADF, VASP, AMBER, figure 2).

2) Le calcul intensif distribué (ou multi-paramètres), pour lequel les grilles de calcul sont exploitées pour partager un grand nombre de tâches, peu ou pas du tout couplées. Comme dans le cas du calcul intensif parallèle, il peut s'agir de résoudre un problème unique, mais ce dernier est composé de plusieurs calculs similaires (surface d'énergie potentielle, famille de molécules, trajectoires de dynamique).

3) Le calcul à la demande, permettant une utilisation temporaire de ressources dont la possession personnelle ne serait pas rentable en termes de capacités de calcul, de logiciels ou de type de matériel par exemple.

Ces différents intérêts soulignent le besoin de grilles de production en Chimie. Par contre, le traitement intensif des données, permettant de générer de nouvelles informations à partir de données géographiquement distribuées, ne représente pas aujourd'hui un intérêt majeur en Chimie.

L'hétérogénéité éventuelle de l'architecture de la grille ajoute un degré de liberté supplémentaire (mais également en contrepartie des difficultés informatiques). En effet, une architecture informatique bien adaptée à une méthode de calcul donnée (par exemple quantique) ne l'est pas forcément pour une autre méthode (par exemple pour une simulation de dynamique classique). Une grille flexible, caractérisée par une hétérogénéité de fonction, permet en outre de répondre à une plus large demande de calcul scientifique, tout particulièrement en ce qui concerne les équipes utilisant différents codes et méthodes.

Enfin, soulignons qu'à une grille de calcul peut correspondre un projet scientifique entre divers laboratoires c'est-à-dire à une grille de compétences et/ou de savoirs, privilégiant la mise en place d'interactions entre individus

3. Utilisation des grilles de calcul en France et en Europe

À ce jour, la communauté française de Chimie n'est pas une grande utilisatrice de grilles, si l'on omet les grilles régionales, comme celles présentes à Grenoble ou Lyon. Néanmoins, dans tous ces cas, la Chimie est souvent minoritaire, tant en terme de moyens qu'en terme de nombre d'utilisateurs. De plus, ces grilles n'ont généralement pas une vocation nationale.

S'il est très difficile d'identifier les causes de cette situation, l'absence d'informations voire la méconnaissance de cette technologie (voir le sondage ci-dessous) peut l'expliquer en partie.

À l'échelle européenne, la situation française reste spécifique. En effet, parmi les cinq organisations virtuelles (VO) en « Computational Chemistry » enregistrées dans EGEE, représentant un total de 322 utilisateurs, la VO dénommée CompChem est sûrement la plus active, pouvant ainsi être prise en exemple. Cette VO regroupe 66 inscrits (avec seulement trois chercheurs français) et représente le troisième exploitant des ressources d'EGEE, après les VO de Bio-Santé et de Physique des Hautes Températures.

4. État de la communauté : le sondage

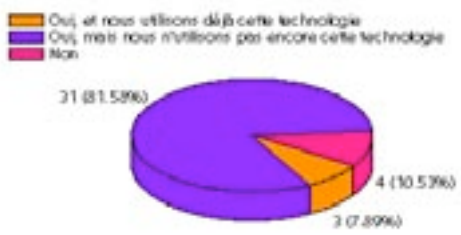


Figure 4 : Question 9 : verriez-vous un intérêt à l'utilisation d'une grille de production ?

Un questionnaire a été publié début septembre sur la liste de diffusion des chimistes théoriciens français (liste du colloque RCTF 2008), afin d'établir l'impact des grilles au sein de la communauté. Ce sondage, composé de 14 questions, visait à déterminer à la fois l'utilisation et la nécessité de moyens de calculs, et à évaluer la connaissance du concept de grille de production. Une part significative des chercheurs contactés y a répondu, 93% d'entre eux ayant un besoin de calcul considéré comme important voire très important (entre 10GFlop et 1TFlop). À ce fort besoin en moyens de calcul correspond cependant un net manque d'informations, puisque 65 % considèrent leur connaissance des grilles comme faible, 15 % la qualifiant de nulle. Seuls 8% des chercheurs utilisent déjà des grilles, mais 82% y voient un intérêt. Enfin, 90% des personnes ayant répondu sont prêts à assister à une formation sur les grilles.

Ainsi, si les chercheurs perçoivent les grilles de calcul comme un bon moyen de subvenir à un manque (chronique) de ressources de calcul, ces dernières représentent également pour eux la possibilité d'envisager des problèmes scientifiques plus complexes et ambitieux que ceux habituellement étudiés avec les ressources de calcul dont ils disposent. D'où l'intérêt de la communauté à s'investir dans le développement d'une grille de production si celle-ci peut répondre à leurs besoins. La majorité des chercheurs ayant répondu au questionnaire souhaitent en outre utiliser la grille pour des applications parallélisées ou utilisant des logiciels sous licence. Enfin, les résultats du sondage ont bien mis en évidence que la communauté de chimistes en France perçoit mal le concept de grille et est très mal informée des diverses possibilités qu'offre une grille de calcul.

5. Prospectives et actions à prévoir pour la Chimie

La faible connaissance des grilles au sein de la communauté ne permet pas la mise en œuvre immédiate d'une démarche directe mettant en œuvre des actions d'information et de formation. Afin de mettre en place une action efficace, il apparaît tout d'abord nécessaire de créer un axe « Chimie » au sein d'une grille existante, puis d'initier une action de diffusion progressive sur son utilisation. Cette dernière sera effectuée en plusieurs étapes :

1) Création d'un point de référence pour la Chimie,
2) Action de sensibilisation et d'information,
3) Pérennisation du point de référence,
4) Actions de formation (continue).

5.1. Création d'un point de référence pour la Chimie

Afin de répandre l'utilisation de grilles dans une communauté « vierge » comme celle de la Chimie, la meilleure stratégie consiste à agir en deux temps : une phase pilote suivie d'une phase d'extension ou phase opérationnelle. La phase pilote doit servir à créer un groupe de travail (chercheurs et techniciens) chargé de la création d'une grille de production en Chimie, délocalisée sur plusieurs sites nationaux ou à déployer des nœuds au sein d'une grille nationale déjà existante. Ce groupe de travail se chargera tout particulièrement d'identifier les problèmes liés aux applications numériques en Chimie, du portage de logiciels de Chimie sur la grille, du déploiement d'un nœud et de l'insertion dans le contexte européen (par exemple VOs Chimie dans EGEE). Cette démarche permettra l'identification d'un ensemble chercheurs-équipement-centres de recherches (et matériel hardware associé) spécialisé en Chimie et constituant un noyau de référence pour la création d'un réseau plus vaste qui sera déployé dans la seconde phase d'extension.

Une fois ces objectifs atteints, le groupe se focalisera sur des applications « phares » de la grille de production en Chimie, afin de bien montrer l'apport de cette technologie à la résolution de problèmes chimiques. La durée nécessaire à la réalisation des objectifs de la phase pilote peut d'ores et déjà être estimée à 24 mois environ.

La phase d'extension commencera une fois que la phase pilote aura pleinement atteint les objectifs escomptés, et sera progressivement mise en place. Elle devra donc répondre aux besoins exprimés et tenir compte des contraintes du moment.

5.2. Gouvernance

Le groupe qui a travaillé sur cette prospective regroupe désormais un nombre significatif de personnes représentatives scientifiquement ou géographiquement de la communauté française des théoriciens en Chimie. Il nous semble donc naturel que ce même groupe soit responsable de la gouvernance pour la partie Chimie et qu'il constitue le noyau du groupe de travail chargé de la première phase du projet.

5.3. Information et formation

Comme mentionné précédemment, l'activité du groupe de travail doit mettre en œuvre une politique globale de formation destinée à la fois aux gestionnaires du système présents dans chaque unité participante (labo ou équipe) et aux utilisateurs. En particulier, lors de la phase pilote une action de sensibilisation à l'usage des grilles (grâce aux exemples fournis par le groupe de travail sur les applications spécifiques en Chimie) sera effectuée au niveau national dans le cadre de séminaires dans des laboratoires et de conférences dans les congrès ou colloques.

La phase pilote se terminera par une première formation nationale « Grille de production en Chimie », centrée sur l'utilisation de grilles de production avec des exemples spécifiques en Chimie. Cette formation sera éventuellement reconduite comme formation continue destinée aux scientifiques et aux ingénieurs des laboratoires de Chimie (1 formation par an).

5.4. Moyens

La diffusion de l'utilisation de grilles en Chimie ne peut pas être limitée à des opérations de formation de chercheurs et techniciens déjà présents au sein des laboratoires. En effet, le succès de cette diffusion dépendra fortement de l'existence d'un personnel hautement spécialisé dans les applications des grilles en Chimie, permettant ainsi un appui logistique permanent. Dans la première phase du projet (24 mois), deux ingénieurs en CDD, chargés de monter et de gérer le projet pilote, pourront être prévus. Ceux-ci auront également en charge la liaison entre les différents laboratoires et équipes impliqués dans le projet. Dans la phase opérationnelle, une configuration minimale de l'équipe sera de deux ingénieurs et d'un chercheur permanent.

Dans le même temps, il faudra prévoir un financement destiné au déploiement d'au moins un nœud de grille de façon à disposer d'un point de référence « hardware » sur lequel les utilisateurs pourront accéder à la grille.

Enfin, des frais de missions en France et à l'étranger (Europe) seront à prévoir pour la participation à des réunions prévues dans le cadre du déploiement de la grille et à son intégration dans EGEE par exemple.

Robert Vergnieux, Stéphane Pouyllau, Bernard Rapacchi

Préalable

La structuration de la numérisation des données scientifiques et de leur pérennisation n'est qu'à ses débuts en Sciences Humaines et Sociales (SHS)SHS. Si l'idée de stockage pérenne semble se faire jour, l'idée que cela puisse être mutualisé n'est pas clairement perçue par l'ensemble de la communauté scientifique. Le potentiel offert par l'accès aux Grilles n'est pas encore identifié par les chercheurs des SHS, cela même si les projets SHS auront, sans aucun doute possible la nécessité de recourir assez rapidement à l'usage des grilles de données et de calculs. Cependant, à l'heure actuelle, peu de porteur de projet en est conscient.

Côté incitation, le Ministère de la Culture ouvre depuis plusieurs années des appels à projet pour la numérisation de « collections » alors que ceux de l'ANR appellent à l'élaboration de corpus. Dans un esprit de maîtrise de la gestion des collections et d'en ouvrir l'accès à tous le Ministère de la Culture semble être le plus en avance. Cependant les chercheurs en SHS développent le plus souvent des bases de données construites sur des axes de recherche et non pas sur l'exhaustivité de collections. Ceci est l'une des différences notoires entre les corpus et les inventaires, non sans incidence sur leur numérisation. C'est à dire que de très nombreuses bases de données en SHS voient le jour sur des supports, des formats et des conditions d'accès des plus hétéroclites, ceci bien que l'usage de normes reconnues soit de mise.

1. Structuration actuelle : le rôle du TGE Adonis pour la numérisation en SHS

Un travail préparatoire a déjà été mené sur la pérennisation des données en SHS, travail compatible avec l'usage des grilles et la mise en place du TGIR. L'effort a consisté à regrouper, dans le cadre d'écoles thématiques, différents acteurs de la numérisation en SHS ayant un rôle moteur. Il est à noter qu'à ces réunions étaient également représentés des centres de calculs opérateurs pour les grilles (CC-IN2P3, CINES). Une réflexion assez large a été ainsi menée sur l'interopérabilité des données et de la mise en place opérationnelle d'un développement à la fois réaliste et ambitieux. La piste suivie la plus pertinente semble être la structuration en Centres de Ressources Numériques spécialisés en termes de contenu pour les SHS (BdD, sons, images, films, SIG, Scan 3D, modèles 3D, etc). Ces Centres de Ressources Numériques, qui fonctionnent en réseaux depuis 2006, sont amenés à servir d'interfaces entre les usagers issus des SHS et les centres serveurs opérateurs sur les Grilles.

2. Le rôle des Centres de Ressources Numériques

Ces Centres de Ressources Numériques doivent s'attacher, dans le cadre du TGE Adonis à promouvoir l'usage de méta-données normalisées (standard internationaux), à la mise en place d'un système d'archivage pérenne basé sur le modèle OAIS (en cours d'expérimentation sur les données sonores) ainsi qu'à la mise en œuvre de l'insertion des données numériques ou numérisées dans des serveurs pérennes et des systèmes d'archivage que seules les grilles de calcul peuvent actuellement proposer au SHS. On constate actuellement qu'un effort est mené en SHS devant aboutir à une structuration des données numériques totalement compatible non seulement avec l'usage des grilles mais encore devant nécessairement s'insérer dans ce genre de structures opérationnelle. Les ressources humaines réparties dans les laboratoires SHS n'ont ni les compétences ni les moyens techniques pour mettre en place et assurer la maintenance de tels dispositifs. En cela il est incontournable que les données numériques soient confiées aux Centres de Calcul via l'intermédiaire des centres de ressources numériques qui catalysent les données tout en les structurant puisqu'ils ont la connaissance des domaines scientifiques concernés.

3. Remarques sur la nature des données numériques en SHS

La manipulation des données numériques sur lesquelles s'appuient les projets de recherche en SHS est en effet plus fortement liée aux chercheurs que dans les autres disciplines. La nature des données traitées est souvent fort éloignée d'une information strictement documentaire. Dans le cas de l'archéologie, par exemple, les données factuelles sur lesquelles vont se construire des projets de recherche reposent en premier lieu sur l'informatisation d'« objets archéologiques inventoriables » qui ne sont ni des collections de musée, ni des livres, ni

des cartes ni des articles de revues. Document unique parfois même disparu mais dont l'existence est connue soit par la conservation de l'objet (une stèle égyptienne par exemple) soit par l'attestation de son existence pouvant être issue de données textuelles (texte latin décrivant la fontaine aux dauphins du Circus Maximus de Rome) ou bien encore objet connu par le biais d'un témoin négatif (tranchées de fondation d'un mur, dont les blocs de calcaire ont été prélevés par des chaufourniers, mais ces tranchées vides attestent ainsi de l'existence d'un mur devenu fantôme). On retrouve ce type de données dans le domaine de l'histoire des sciences avec les manuscrits et les cahiers de laboratoires des savants et dans les correspondances scientifiques qui permettent, en négatif, de comprendre les réseaux d'échange des connaissances scientifiques. Le document source par sa description dans une base de données devient alors une entité numérique. Cette source va à son tour fournir une série de données, ouvrant alors sur de larges jeux de connaissances. Le texte inscrit sur une stèle livre des informations multiples au-delà du seul objet que constitue la stèle qui le porte. Ces informations complexes ne peuvent appartenir à un chercheur en particulier, elles appartiennent à la collectivité tant les pistes de recherche peuvent être multiples à partir d'une même source.

Cependant c'est le travail particulier de tel ou tel projet de recherche qui va opérer la numérisation de ces données factuelles. Afin d'éviter que chacun dans son coin ne développe ce qui a déjà été fait par un autre, les bases de données doivent être conçues inter opérables dès l'origine et cela au niveau même des données factuelles.

4. Un gros verrou !

La notion même de « grille » est pour ainsi dire totalement inconnue en SHS.

Bon nombre de chercheurs « bricolent » encore leur propre base de données à partir de logiciel tel que FikeMakerPro. Le fait de présenter les grilles comme une véritable solution d'avenir pour l'archivage et l'accessibilité peut éventuellement éveiller l'attention de la communauté. Les pratiques en tant que grille de calcul seront encore plus complexes à faire valoir. Ces verrous tendent à se lever avec l'apparition du travail en mode projet et les contraintes imposées par les appels à projet sur la qualité, la pérennité et l'interopérabilité des bases de données « finançables » et la professionnalisation des métiers d'accompagnement de la recherche en SHS principalement ceux des humanités numériques ou digital humanities dont tous les acteurs nationaux sont aujourd'hui structurés autour du TGE Adonis (http://www.tge-adonis.fr), reconnu depuis la roadmap du Ministère comme l'une des TGIR des SHS.

5. Grilles de données et SHS

Le souci, pour les porteurs de projet, de devoir rendre leurs bases de données accessibles via le web, comme le préconisent souvent les appels d'offre, incite ces chercheurs à confier leurs données à des fins d'accessibilité web. Il n'est pas réaliste de laisser tout porteur de projet en SHS dialoguer en direct avec les centres de calcul ou les opérateurs de la grille pour que ces derniers prennent en charge leurs données numériques. L'apparition des Centres de Ressources Numériques en SHS garantit une dimension opérationnelle à l'insertion des données numériques sur les réseaux et permet au passage de les normaliser afin de s'assurer de leur inter opérabilité et de leur cohérence. Ainsi les opérateurs des grilles n'auront, par type de données, à travailler qu'avec une seule équipe (CRN) servant de point d'entrée aux autres.

6. Grilles de calcul et SHS

Cet aspect a encore été peu envisagé du côté de la structuration des données numériques en SHS cependant il est important de prendre en compte dès aujourd'hui les besoins identifiables dans la mesure où, une fois apparus, ils seront très gourmands en temps de calcul. En repartant de l'exemple de l'archéologie, le stockage et la manipulation de données 3D sont actuellement un véritable problème. En effet comment explorer les données visuelles contenues dans des fichiers de tomographie de quatre vingt giga-octets, comment se déplacer en temps réel dans des modèles 3D excessivement détaillés. Il semble que la mise en place sur la grille d'outils performants de manipulation d'objets 3D temps réel soit à prendre en compte. Effectuer des calculs de rendu de films en images de synthèse toujours à partir de modèle 3D de l'archéologie, peut atteindre pour une production de dix minutes, un temps de calcul de 156 jours CPU. La délocalisation de tels calculs sur la Grille prend alors toute sa pertinence. Des tests de faisabilité sont actuellement pratiqués par la PFT3D de Bordeaux (Archéogrid - Archéovision) dont les calculs de rendu d'image sont effectués expérimentalement sur les clusters du CC-IN2P3. Ces développements intéresseront rapidement une grande partie de la communauté tant l'image numérique entre au cœur des programmes de recherche. Notons que la mise en place de tels outils pourra servir bien au-delà de la communauté des archéologues puisque la 3D envahit progressivement tous les champs disciplinaires et la nécessité de disposer en ligne de telles ressources sera alors une demande plus fournie. Sur le plan du traitement des données cartographiques/iconographiques 2D, les besoins de calcul de rendu et de tuilage d'images représentant de très grandes surfaces (données cartographiques, satellitaires, etc.)

sont en forte croissance. Le CRN pour les données iconographiques (le Centre National pour la numérisation de sources visuelles – http://www.cn2sv.cnrs.fr) travaille actuelle sur la mise en relation entre des documents cartographiques 2D géo référencés et les interfaces de consultation simple tel que Google Maps ,etc. : les besoins de tuilage d'images de très grande taille (plusieurs Go par images) seront de plus en plus importants et doit être également mise en relation avec le besoins de visualisation des données « à la volée » via des interfaces passant par le web (donc avec des fonctions de dégradation/zoom dynamiques). Les besoins de calcul pour le traitement des données vidéos et sonores sont aussi naissant, il s'agira là aussi de mutualiser les outils de traitement tout en les « distribuant » via des flux auprès des communautés scientifiques.

Figure 1: Production d'une séquence vidéo HD à partir d'un modèle numérique 3D issu d'un programme de recherche (ANR BLANC 2008/ATON3D). Une séquence de 10 minutes entraîne le calcul de rendu de plus de 20 000 images. Le calcul de la séquence nécessite 150 jours CPU.

7. Recommandations

L'intégration de l'usage des grilles pour les SHS représente une opportunité de structuration pour toutes les disciplines qui les composent. Tout est mûr pour une telle aventure mais il reste à convaincre la communauté de la chance qui se présente à elle de sortir des solutions numériques individuelles et non pérennes.

L'action à mettre en place peut se décomposer en quatre phases :

1) Aide à la structuration des acteurs du numérique en SHS

Aider au renforcement des Centres de Ressources Numériques qui se structurent au sein du TGE Adonis. Ils joueront le rôle d'interface entre les porteurs de projets de recherche et l'Institut des Grilles et ses opérateurs. Ainsi chaque CRN connaissant les possibilités opérationnelles ou les développements possibles saura pendre en charge l'aide au portage des données numériques sur les grilles. Ils garantiront l'interopérabilité des données et leur pérennité au regard de la recherche. Cette action aura comme résultante de dériver progressivement les données numériques des chercheurs SHS au sein de structures informatiques moins disparates et dans des réseaux de haut de gamme et non plus dans des solutions souvent système « D » disparaissant avec les départs à la retraite. La recherche en SHS trouvera en cela une plus grande cohérence et la mise en place effective de silos de « connaissances » spécifiques à la discipline. Les financements de l'État trouveront dans cette démarche une meilleure visibilité sur la production de données « numérisées » et mutualisées. Cela limitera de fait la déperdition et facilitera l'interopérabilité.

2) Aide à l'émergence de projets « exemplaires »

Aider deux à trois projets pouvant servir d'exemples pédagogiques sur le potentiel de l'usage des grilles. Des projets de recherche s'appuyant des base de données images de très haute définition, des calcul de films en images de synthèse HD, sur de la manipulation intensive de flux vidéo ou d'enregistrements sonores peuvent donner naissance à des cas d'école pour une meilleur prise de conscience du potentiel des grilles.

3) Aide à la diffusion de ces pratiques

L'organisation d'une école thématique autour de ces usages pourra se faire une fois les premiers exemples validés. La diffusion de cahier de recommandations aux équipes de recherches fait partie de la sensibilisation de la communauté.

4) Aide par la mise en place d'appel à projets en SHS faisant l'usage des grilles

Cette étape ne peut être atteinte que dans la mesure où les étapes antérieures l'ont également été.

Listes des Centres de Ressources Numériques en cours de structuration en tant que composantes du TGE Adonis devenu depuis la roadmap du Ministère de début janvier 2009 l'une des quatre TGIR du secteur SHS :

- « CRDO » Centre de ressources pour la description de l'oral, géré par le Laboratoire Parole et Langage – LPL (UMR 6057) et le Laboratoire de Langues et civilisation à tradition orale – Lacito (UMR 7107). Adresse web :http://www.crdo.fr

- « CNRTL » Centre National de Ressources Textuelles et Lexicales,géré par le laboratoire d'Analyse et traitement informatique de la langue française – ATILF (UMR 7118) et le Centre d'études supérieures de la renaissance – CESR (UMR 6576). Adresse web :http://www.cnrtl.fr

- « TELMA » Traitement ELectronique des Manuscrits et des Archives », géré par l'Institut de recherche et d'histoire des textes – IRHT (UPR 841) et l'École nationale des Chartes. Adresse web : http://www.cn-telma.fr

- « M2IAS » Méthodologies de Modélisation de l'Information Spatiale Appliquées »,géré par le Centre d'Etude de l'Inde et de l'Asie du Sud – CIAS (UMR 8564). Adresse web : http://www.m2isa.fr

- « CN2SV » Centre National pour la Numérisation de Sources Visuelles », géré par le Centre de Recherche en Histoire des Sciences et des Techniques – CRHST (UMR 8560). Adresse web : http://www.cn2sv.fr

- « ARCHEOGRID » Conservatoire National pour les données 3D du Patrimoine, soutenu par le CPER Région Aquitaine et géré par la Plate-Forme Technologhique 3D, Institut Ausonius (UMR 5607). Adresse web : http://www.archeovision.cnrs.fr

- Le TGE ADONIS « accès unifié aux données et documents des sciences humaines et sociales », (UPS 2916). Adresse web : http://www.tge-adonishttp://www.tge-adonis.fr/.

Groupe 6 : Mathématiques-Physique-Fusion

Pierrick Micout, Patrick Leboeuf, Stéphane Cordier, Bernard Guillerminet

1. Mathématiques

En ce qui concerne les mathématiques, il semble que, soit les mathématiciens ne connaissent pas les grilles, soit ils les utilisent sans le savoir.

Si malgré tout on prend le thème « Grille pour les maths » au sens large, on constate que :

- Le GDS Mathrice a impulsé une dynamique d'utilisation des réseaux pour une mutualisation des ressources en jeton de logiciel payant. Mais ces progiciels étant installés directement sur les machines des utilisateurs, on est vraiment loin de la notion de grille de calcul.

- Une autre utilisation des ressources est l'accès à des bases de données bibliographiques en ligne. Les GDS du CNRS « Mathrice » et RNBM sont en pointe pour ces accès à des ressources qui coûtent de plus en plus cher. La documentation en mathématiques est considérée comme un « grand instrument » pour le (futur) institut de maths.

- Il y a aussi plusieurs projets particuliers liés à la cryptographie (cassage de code RSA) ou au projet GIMPS (recherche des plus grands nombres de Mersenne) http://www.mersenne.org/ ou encore les applications des méthodes Monte-Carlo (moyennes sur des réalisations indépendantes de simulations) qui peuvent être distribuées sur la grille.

- on peut aussi noter l'utilisation de la grille dans les logiciels de calcul formel http://www.wolfram.com/products/gridmathematica/ ou dans ceux de calcul scientifique (scilab en utilisant la plateforme proactive) http://proactive;inria.fr/index.php?page=scilab .

- Il y a aussi les projets BOINC (1,2 pétaflops établis) (http://boinc.berkeley.edu/, voire (http://boinc.berkeley.edu/projects.php (catégorie «math and strategic game») et aussi la liste des participants académiques français (peut-être un peu réduite)(http://www.boinc-af.org/content/category/7/109/294/ .

- Cette faible participation du monde académique français semble être générale.

En résumé, la communauté des mathématiciens ne semble pas très impliquée dans l'utilisation des grilles, qu'elles soient de production ou de recherche, pour le moment. Il faut noter cependant qu'un projet ANR (GCPMF) de mathématiques financières existe et semble progresser normalement.

2. Physique

En dehors des communautés liées à l'astrophysique, à la physique nucléaire, à la physique des particules et à celles travaillant sur les nanotechnologies (toutes traitées par d'autres groupes de travail), l'utilisation des grilles de production ne semble pas être effective dans le domaine de la physique.

3. Fusion

Dans le monde de la fusion contrôlée par confinement magnétique, plusieurs personnes réfléchissent à la possibilité d'utiliser les grilles de production et les supercalculateurs qu'ils appartiennent à des grilles comme DEISA 2 ((http://www.deisa.eu/fr/) ou pas. Ils sont engagés dans un projet du 7ème programme cadre de recherche et développement de l'Union Européenne (projet Euforia : EU fusion for Iter Applications,(http://www.euforia-project.eu/EUFORIA/) et peuvent déjà utiliser la grille EGEE ((http://www.eu-egee.org/). Ils travaillent plus particulièrement sur :

- la convergence entre les intergiciels « gLite » (http://glite.web.cern.ch/glite/) et « Unicore » (http://www.unicore.eu/), • l'accès aux données, • l'interactivité, • la visualisation, • l'ordonnancement des tâches nécessaire à l'analyse des données et à la simulation.

Il faudrait améliorer l'information sur les grilles de production auprès des autres utilisateurs potentiels (optique, physique du solide, etc.).

Groupe transverse 1 : Grilles de données

Françoise Genova, Dominique Boutigny, Vincent Breton, Sébastien Denvil, Michel Kern, Pierrick Micout

1. INTRODUCTION

De plus en plus de disciplines scientifiques prennent conscience du fait qu'il est important de pouvoir conserver et réutiliser les données qu'elles produisent – données d'observations, données dérivées, données de modélisation. Le partage de l'information est un élément essentiel qui sous-tend le développement des e-Infrastructures. Plusieurs projets de ce type figurent parmi les Très Grandes Infrastructures de Recherche identifiées au niveau national et dans la feuille de route européenne ESFRI, et on voit apparaître dans le Septième Programme Cadre des thématiques centrées sur les « Scientific Digital Repositories ».

La problématique des données a plusieurs aspects : stockage, préservation sur le long terme, distribution à des communautés plus ou moins larges, élaboration de services pour faciliter l'utilisation des données. Certains aspects sont techniques et peuvent trouver des solutions génériques, par exemple la conservation physique des données, qui suppose des changements réguliers de support. Il est aussi nécessaire que les disciplines se mobilisent pour identifier la manière dont les données seront utilisées, pour définir un cadre adéquat pour la description des données et le développement des services d'accès. Enfin, les producteurs de données doivent accepter de fournir un effort important de description de leurs données pour que celles-ci puissent être réutilisées par d'autres.

La communauté Grille développe de plus en plus d'applications qui nécessitent, non seulement du calcul intensif et du stockage massif de données, mais aussi d'accéder à des bases de données préexistantes, hétérogènes, et dont les opérateurs sont indépendants. Elle étudie donc l'inclusion des bases de données et de leurs technologies dans l'environnement de la Grille. Cet aspect est techniquement moins mûr que l'utilisation traditionnelle de la Grille, et il faudra poursuivre le dialogue entre les besoins et les développements au-delà de cet exercice de prospective. Réciproquement, on peut dire qu'un ensemble de données distribuées, auxquelles on peut accéder de façon transparente (sans avoir à apprendre dans chaque cas les spécificités d'une interface particulière) via des services qui permettent de les comparer et de les combiner, est une « grille de données et de services », mais qui ne fait pas nécessairement appel aux techniques développées pour les Grilles de Production. Il n'y a pas en effet de solution unique aux différents aspects du problème.

Dans ce contexte, le « Groupe transverse Grilles de Données » s'est attaché à rassembler des exemples provenant de disciplines différentes. Les besoins, la culture et le contexte de travail de chaque discipline orientent le choix des solutions, et il était important d'identifier l'apport effectif ou potentiel de la Grille et de ses techniques. Les quelques exemples ci-dessous illustrent la diversité du paysage.

2. ÉTUDES DE CAS

2.1. Santé publique

C'est un domaine d'application privilégié des technologies Grille, en particulier à cause des exigences particulièrement fortes en terme de contrôle et d'accès aux données. On peut citer par exemple deux projets en Auvergne qui étudient l'utilisation des grilles pour la mise en place d'un réseau sentinelle régional sur le cancer et international sur la grippe aviaire. Ces projets s'appuient sur le savoir-faire des laboratoires auvergnats dans l'utilisation des grilles pour la santé et les sciences du vivant, illustré dans les initiatives HealthGrid et WISDOM, et sur l'infrastructure de grille régionale AuverGrid. Pour ce type d'applications, la Grille permet de créer un réseau de bases de données fédérées, où il est possible d'aller chercher des informations. Les données sont stockées sur le lieu où elles sont produites (hôpital, cabinet de radiologie, laboratoire d'anatomopathologie, …), et l'exécution de requêtes est conditionnée à un droit d'accès dont les limites sont fixées par le propriétaire des données. Ainsi, il n'est plus nécessaire de saisir à nouveau les données, source d'erreur, ni de les dupliquer, ce qui constitue une faille de sécurité. Elles restent sur leur site de génération, sous le contrôle des professionnels de santé qui ont supervisé leur production, et elles sont rendues visibles à des acteurs dûment accrédités pour des requêtes préalablement définies et validées par l'ensemble des partenaires du réseau.

Dans le cas de menaces globales de type grippe aviaire, la valeur ajoutée de la grille vient notamment de sa capacité à mettre en commun des informations au niveau mondial, tout en les laissant complètement sous le contrôle des institutions qui les produisent (ministères, établissements hospitaliers, instituts de recherche). L'objectif est donc d'améliorer la réponse globale aux maladies émergentes en améliorant la collecte, la mise à disposition et l'accès aux données concernant la maladie à des fins de recherche et d'alerte. La grille permet aussi dès aujourd'hui de mobiliser à la demande des ressources en calcul très importantes en cas d'urgence.

2.2. Modélisation climatique

Les données provenant des modèles climatiques sont produites en grande majorité sur les centres de calculs nationaux tels l'IDRIS ou le CCRT, voire internationaux dans certains cas, comme le Earth Simulator japonais. Plusieurs initiatives internationales développent des standards pour décrire les modèles climatiques et les données produites, en particulier le projet européen METAFOR et le projet américain CURATOR, ouvrant la voie à une généralisation de l'approche par base de données des résultats de simulations climatiques. La quantité de données générées devenant conséquente, l'idée de base est d'éviter au maximum de les déplacer, de les laisser dans les centres de calcul, et de leur donner un accès unifié en construisant une « grille de données ». Celle-ci doit impérativement inclure les centres de calcul nationaux (IDRIS, CCRT, CINES, et quelques mésocentres).

Le prochain rapport de l'International Panel on Climate Change (le cinquième rapport : IPCC AR5) s'appuiera effectivement sur des données distribuées, et non plus centralisées. Dans ce contexte, l'Institut Pierre Simon Laplace étudie de près les potentialités de la grille et son apport possible en terme de gestion de données. Il vise à intégrer l'ensemble de l'architecture (data nodes, gateways, global services), avec la mise en place de data nodes dès 2009. L'apport des fonctionnalités des portails devrait permettre d'élargir de façon significative l'éventail des utilisateurs des données de simulation numérique.

2.3. Fusion/ITER

Les simulations sont effectuées sur des clusters, sur EGEE ou sur des supercalculateurs (par exemple ceux des centres nationaux), éventuellement connectés par l'intermédiaire de DEISA. L'objectif des projets EUFORIA (EU for ITER applications) et ITM (Integrated Tokamak Modelling) est de développer un modèle validé de machine de fusion incluant le plasma. Les codes accèdent aux données de la même manière quelle que soit leur localisation, et les codes exécutés sur la grille ou sur les calculateurs à haute performance doivent pouvoir accéder à des données stockées en dehors de l'infrastructure grille. Ces échanges de données sont fréquents et peuvent être importants. Les enchaînements de tâches (workflow) sont basés sur les outils multidisciplinaires KEPLER et PTOLEMY-II, développés respectivement par le Supercomputer Center University of California à San Diego et par l'University of California à Berkeley.

2.4. Astronomie

La discipline astronomie a une longue tradition de partage des données, et l'essentiel des observations obtenues par les grands télescopes sol et spatiaux est mis à disposition de la communauté dans des centres de données distribués tout autour du monde. Le concept d'Observatoire Virtuel, qui a émergé au tournant du siècle, s'est rapidement développé ces dernières années : son objectif est de donner accès à l'ensemble des données disponibles, de façon transparente pour les utilisateurs, avec des outils pour visualiser, combiner et analyser les données. Les standards d'interopérabilité (description des données, protocoles d'accès, langage de requête, liste des ressources, …) sont définis au niveau international par l'International Virtual Observatory Alliance, qui regroupe une quinzaine de projets nationaux (y compris le projet européen Euro-VO et le projet français). Les standards essentiels sont disponibles, utilisés par de nombreuses archives, et des portails d'accès à l'information ont été développés. Le projet passe progressivement en phase opérationnelle, et le projet EuroVO-AIDA (Astronomical Infrastructure for Data Access), coordonné par le CNRS, organise cette transition au niveau européen.

Les interactions avec la Grille de Calcul sont explorées dans le cadre d'un Groupe de Recherche de l'Open Grid Forum et par le projet européen EuroVO-DCA (Data Centre Alliance). Certaines techniques développées par les projets Grille, en particulier pour la sécurisation de l'accès aux données, pourraient être réutilisées. Réciproquement, l'Open Grid Forum s'intéresse à certains standards développés par l'IVOA.

2.5. Physique des hautes énergies

Les expériences de physique des hautes énergies et en particulier celles installées sur le LHC au CERN produisent des quantités très importantes de données qui doivent pouvoir être accédées par toutes les équipes de recherche distribuées mondialement. Les expériences du LHC ont mis en place une architecture de grille (LHC Computing Grid ou LCG) basée sur l'intergiciel EGEE en Europe, OSG aux États-Unis et ARC dans les pays nordiques. La LCG repose sur une architecture hiérarchique à 4 niveaux (le CERN est le niveau 0, les stations des utilisateurs finaux le niveau 3). Le rôle de cette grille est double puisqu'en plus de permettre le traitement des données, elle fournit l'architecture nécessaire au stockage pérenne des données, à leur catalogage, à leur distribution et à leur sécurisation.

Une fois encore, l'accès aux données pour les scientifiques sera totalement transparent. Pour un projet de cette envergure, seule une architecture distribuée permet d'obtenir la puissance de traitement nécessaire. De plus, la nature des données s'y prête: on doit réaliser un très grand nombre de traitements indépendants.

3. CONCLUSIONS

Le point de départ, pour toutes ces applications, est l'évolution générale vers la production et le stockage des données de façon distribuée, parfois près des lieux de production, mais pas toujours. Certaines applications utiliseront la Grille stricto sensu, d'autres certains éléments de la Grille. Ces quelques exemples illustrent la diversité des besoins, parmi lesquels on voit apparaître en particulier :

- L'accès aux données à partir de programmes exécutés sur la grille,
- La sécurisation des accès, impérative dans certaines disciplines, et son corollaire, l'authentification des utilisateurs,
- La définition de standards pour assurer l'interopérabilité.

Le développement d'outils – génériques, ou adaptés au contexte particulier d'une ou plusieurs applications - pour permettre la recherche, et l'indexation, dans un grand volume de données. Il s'agit d'un sujet de recherche en informatique, pour lesquels les interlocuteurs pertinents doivent être identifiés. Il faudrait certainement également encourager les échanges d'expériences dans ce domaine.

Des logiciels tels que le Storage Resource Broker (SRB) ou son successeur iRODS développés à San Diego (SDSC) sont des systèmes de grilles de données qui permettent de transférer et de cataloguer des données de manière simple. Il est également possible de définir des services associés aux données qui seront automatiquement appliqués en fonction de leurs provenances ou de leurs caractéristiques. SRB / iRODS est utilisé, entre autre, par la communauté biomédicale ; le réseau BIRN (Biomedical Informatics Research Network) aux USA est un exemple de déploiement à grande échelle d'une grille de données basée sur SRB.

Il faudrait, dans ce domaine en émergence, pérenniser ce groupe de réflexion. Il serait utile de mettre en place un point de rencontre régulier pour favoriser les interactions entre les équipes qui développent des applications et les spécialistes de la Grille, et partager les expériences dans les différents domaines. Il faudrait également disposer d'aide à l'implémentation et de tutoriels en ligne pour aider les débutants à s'approprier les technologies de la Grille.

Groupe transverse 2 : Grilles régionales - relations production - GRID5000

Frédéric Desprez, Laurent Desbat, Vincent Breton, Marie-Pierre Delest, Michel Jouvin, Jean-Pierre Meyer, Olivier Richard, Michel Kern, Catherine Rivière, Emmanuel Jeannot, Pierre-Louis Reichstadt, Isabelle Cuzor, Guy Wormser

Thierry Priol (Grille recherche) et Pierre Valiron (Observatoire de Grenoble, décédé le 31 août 2008) ont fortement contribué au groupe de travail.

1. SYNTHÈSE DU TRAVAIL SOUS FORME DE RECOMMANDATIONS

1. La synergie entre la « grille recherche » (GRID5000/Aladdin) et la « grille de production » est évidente. Les recherches développées au sein de la grille recherche permettent d'expérimenter de nouvelles technologies de grille et donc de préparer les évolutions de la grille de production. Inversement, les besoins en matière de production peuvent induire des recherches spécifiques en matière de grille. Les contraintes et les objectifs d'utilisation de la grille recherche sont très différents de celles de la grille de production. Les techniques de virtualisation ne sont pas suffisamment mûres aujourd'hui pour une utilisation de la grille de recherche pour la production. La séparation des deux architectures est nécessaire et utile. Les deux axes « grille recherche » et « grille de production » doivent donc être conjointement soutenus.

   Il faut de même soutenir les groupes de travail communs « grille de production »/ »grille recherche » et les séminaires communs consacrés aux expériences, aux avancées et aux limites des technologies de grilles, afin de maintenir une cohésion forte entre les deux axes production et recherche.

2. Les mésocentres[13] ont pour vocation d'offrir des moyens de calcul intensif, de traitement et de stockage de données, des services associés et donc de développer l'expertise locale de la modélisation numérique et du calcul. Leur activité est principalement centrée sur la production locale et intermédiaire, l'accompagnement de projets scientifiques vers le calcul intensif et la formation. Ils servent de tremplin vers le centres de calcul nationaux. Cette activité est essentielle et doit être maintenue en particulier pour encourager les chercheurs à « penser PétaFlops ». Leur mise en réseau devrait être organisée au niveau national (demandée par la CPU[14] avec le soutien de GENCI[15]) pour permettre une interaction structurée avec les centres de calcul nationaux d'une part et les grilles recherche et production d'autre part.

3. Les activités de grilles doivent se développer en lien avec les mésocentres (en particulier, dans ceux qui en expriment le besoin). Certes, certains problèmes de modélisations numériques ne relèvent pas des grilles de calcul mais, dans la mesure où le coût de traitement dans une grille devrait-être moindre que dans un centre national pétaflopique, il faut encourager l'utilisation des grilles dans les mésocentres pour tous les problèmes qui en relèvent. Enfin et surtout, certaines communautés (climat, géophysique, astrophysique, biologie, etc.), partagent des données et des codes : la mise en grille de leurs mésocentres au niveau international (ou de leurs activités dans une grille de mésocentres) devrait renforcer le développement scientifique de ces communautés et devrait être soutenu.

   3.1. Les mésocentres peuvent héberger des nœuds de grilles.

   - Avantages : utilisation de l'infrastructure existante, permet de développer l'expertise des ingénieurs locaux et donc favoriser la diffusion des technologies de grille, dans toutes les communautés scientifiques.

   - Limites : les ingénieurs des mésocentres sont déjà bien chargés => le coût de l'administration des nœuds locaux de grille doit être minimal sinon les équipes d'ingénieurs locaux doivent être renforcées. Il convient en particulier de mettre en œuvre des outils de grille standard dont l'administration doit être la plus légère possible.

   3.2. Les utilisateurs doivent être formés à l'utilisation des grilles (introduction aux technologies de grilles, accompagnement de projets scientifiques à la mise en grilles de leurs programmes). Il faut renforcer l'activité de formation et de soutien local aux projets scientifiques qui ont besoin des grilles. Les mésocentres pourraient être utilisées en relais de formation (cf. état des lieux des formations dispensées dans les mésocentres[16]).

   3.3. L'expérimentation des techniques desktop grids, ou toute autre technologie (cloud computing etc.), pour intégrer ponctuellement des ressources de mésocentres doit être soutenue.

4. Les grilles régionales devraient s'intégrer naturellement dans le paysage national ou international des grilles, soit en participant à une grille internationale de production (exemple de GRIF[17] ou d'Auvergrid[18]), soit en développant des liens étroits avec la grille recherche (exemple deCiGri[19]).

2. LES RÉSULTATS PLUS DÉTAILLÉS DU TRAVAIL DU GROUPE

L'objectif du Groupe est d'étudier les grilles régionales, leur intérêt, leur relation avec les grilles nationales et internationales, leur management, leur financement, leur typologie. Dans ce même groupe, la question de la meilleure synergie entre les nœuds grille 5000 et les nœuds locaux d'une grille nationale sera à aborder.

Dans le contexte de la prospective nationale sur les grilles, nous abordons ici la notion de grille régionale avec pour objectif de proposer des modes de collaboration et partage (de mise en grille) entre ces grilles régionales et une infrastructure nationale de grille. Nous abordons aussi la synergie entre Aladdin, IDGrilles et les grilles régionales.

*2.1. Synergie entre grilles régionales, mésocentres et grille nationale de production (*IdGrilles[20]*)*

Les initiatives grilles régionales que nous avons recensées sont liées aux mésocentres de calcul (voir http://calcul.math.cnrs.fr/ en particulier http://calcul.math.cnrs.fr/spip.php?rubrique5).
• Le mésocentre Auvergne[21] en particulier la grille AuverGrid répartie sur douze nœuds en région Auvergne.
• Le mésocentre Ile-de-France[22] en particulier la grille le GRIF (Grille de production pour la recherche en Ile-de-France) répartie sur 6 sites de la région Ile-de-France.
• Le mésocentre CIRA (Calcule Intensif en Rhône-Alpes[23]) en particulier sa grille régionale Rhône Alpes Grid, RAGrid[24] : RAGrid s'appuye CiGri (CIMENT GRID), DIET[25] et EGEE[26] . RAGrid est répartie sur une dizaine de nœuds en région Rhône-Alpes et exploite des machines dédiées à l'enseignement grâce à des techniques de desktop grid.
L'intérêt de ces grilles régionales est de fédérer des moyens de calcul souvent distribués dans des sites différents afin de pouvoir partager des ressources. Au delà, il s'agit bien souvent d'un partage d'expertises techniques d'ingénieurs mais aussi scientifiques (formation, collaborations, séminaires, sont souvent associés à ces grilles). Elles sont l'occasion de collaborations entre les experts de l'informatique distribuée et les utilisateurs, pour une utilisation plus efficace des moyens de calcul mais aussi pour des spécifications au plus proche des besoins des utilisateurs, des intergiciels et des outils logiciels d'interface aux moyens de calcul et de grille ainsi que des services aux utilisateurs.
Enfin, ces grilles régionales ont déjà un lien fort avec les initiatives nationales de grille : IDG et EGEE pour AuverGrid et GRIF, GRID 5000 et Aladdin pour RAGrid (via CiGri et DIET) mais aussi EGEE via la collaboration entre CIRA et le CC de l'IN2P3 de Lyon et via le projet MUST[27].
Le financement et le management de ces grilles régionales est lié au financement et au management des mésocentres qui les abritent. En général, ces mésocentres sont financés par des CPER, des contrats quadriennaux, les établissements qui les abritent (cet aspect devrait être renforcé dans le contexte de la LRU), des projets ANR, des projets européens, voire des actions scientifiques comme les RTRA.

Ces grilles sont généralement des grilles de calcul. Certaines participent à EGEE. Certaines exploitent la nuit des machines dédiées le jour à l'enseignement. Elles sont en général administrées par un ou des ingénieurs dédiés et leur utilisation est arbitrée (si nécessaire) par un comité de pilotage (ce qui est assez simple lorsque le nombre d'utilisateurs est faible).

Les synergies entre la grille nationale et les grilles régionales peuvent être nombreuses : hébergement commun et partage de matériels, administration communes des matériels et logiciels, diffusion d'expertises via les mésocentres vers les utilisateurs. Les mésocentres pourraient plus largement héberger des nœuds de la grille nationale.

2.1.1. Les enjeux

- Bonne couverture nationale, bonne diffusion des technologies de grille.

- Obtenir la participation des Régions et des universités pour le financement, l'exploitation et l'utilisation des grilles.

La pénétration des mésocentres peut être un atout majeur pour les points 1/ et 2/ (par essence la grille est distribuée : utilisation possible des ressources des mésocentres (humaines, matérielles [bâtiment, équipement, matériel de calcul et de stockage]).

- Meilleure diffusion par une plus grande proximité des technologies de grille.

- Valorisation économique : les grilles régionales permettent de tisser des contacts avec le réseau local des PME/PMI. Elles favorisent aussi l'émergence d'actions de transfert de technologie en partenariat avec les collectivités territoriales.

2.1.2. Les difficultés à surmonter

- Les mésocentres régionaux qui hébergent des grilles ont souvent comme vocation première de servir de tremplin vers les centres de calcul nationaux (former par la pratique leurs utilisateurs à "penser pétaflops" sur des architectures semblables, mais à une échelle inférieure, à celles des centres de calcul nationaux). Cette vocation essentielle doit être soutenue et ne doit pas être mise en danger par le déploiement et l'administration de nœuds de grilles.

- Difficultés technologiques (interopérabilité des différentes grilles). Il est essentiel de mettre en œuvre des technologies standard, pérennes, interopérables, dont le coût d'administration est le plus faible possible.

- Fortes limites en IATOS dans les mésocentres (exprimées lors de la journée dédiée aux mésocentres de calcul[28], du 13 février 2008). Cette limite renforce la nécessité de partage d'expertise et d'une bonne structuration globale.

- Prise en compte des politiques d'utilisation des grilles :

  - Exemple de CiGri : les travaux de grilles CiGri sont déployés sur l'ensemble des machines du CIMENT[29] et sur des machines de la Fédération Lyonnaise de Calcul Haute Performance[30]). Ils sont donc potentiellement en concurrence avec les travaux soumis dans le mésocentre hors grille. Sur les machines de CIMENT, les travaux de grilles sont les moins prioritaires sur le réseau des machines du mésocentre : ils peuvent être à tout moment interrompu.

  - Exemple d'AuverGrid : la grille régionale en Auvergne rassemble 12 grappes de PC dont 6 sont mutualisées à travers l'intergiciel gLite, les autres grappes fonctionnant de façon autonome. Les grappes mutualisées sont principalement exploitées par des utilisateurs sur la grille au niveau régional et à travers les organisations virtuelles d'EGEE. La principale difficulté est d'accompagner l'émergence de l'activité sur la grille au niveau régional tout en optimisant l'exploitation des ressources à travers des organisations virtuelles thématiques d'EGEE.

  - Exemple de GRIF : GRIF est né pour être «exclusivement» un «gros» nœud EGEE en région Ile-de-France, pour faire face en particulier aux besoins des expériences LHC mais ouvert dès le départ à toutes les communautés (20% dans le projet initial, beaucoup plus maintenant). Il a été créé à partir d'un petit embryon de ressources grille existantes en 2004 et est devenu un mésocentre au printemps 2008. GRIF rassemble une équipe technique de 20 personnes réparties sur les 6 sites et représentant 10 ETP, pour la plupart préexistant au projet, et une infrastructure aux ressources très significatives (5000 MSI2K, 500 TB, 1 réseau interne privé et un lien externe 10 Gb/s aujourd'hui) administrée de façon unifiée (concrètement administrable dans son ensemble par n'importe quelle personne de n'importe quel site). Cette mutualisation, sans concentration et sans assèchement d'aucun site, a permis une augmentation considérable de l'expertise grille disponible en région Ile-de-France (il y avait au départ du projet 3 à 4 personnes avec une expertise grille).

La synergie entre les nœuds de grilles régionales et ceux d'une grille nationale/internationale (production ou recherche) existent dans le cas des grilles des mésocentres. Elle pourrait être renforcée. Pour cela, il est important de pouvoir simplement gérer les différentes priorités d'utilisation selon les priorités scientifiques, de même que l'hétérogénéité des besoins et des systèmes. Dans cet objectif, il est essentiel de poursuivre les recherches sur les grilles, en particulier sur les méthodes permettant leur interopérabilité.

2.2. Synergie entre grille de recherche et grille de production

La nature et les objectifs de l'action de développement technologique Aladdin de l'Inria et de l'IDG sont complémentaires. IDGrilles et les grilles régionales ont principalement des objectifs de production pour des utilisateurs qui manipulent les grilles comme un outil au service de leurs recherches en physique (des particules

principalement) mais aussi chimie, sciences de l'univers, biologie, etc. Ils souhaitent en général disposer d'une grille relativement disponible, souple d'utilisation, très fiable et très stable. Par contre, les utilisateurs d'Aladdin utilisent la grille comme un grand instrument de recherche sur les grilles et plus largement sur l'informatique distribuée. Ils implémentent fréquemment de nouveaux systèmes sur les nœuds de la grille, ils peuvent mobiliser à tout moment l'ensemble de la grille recherche pour une expérimentation de nouvelles technologies de grilles. Le partage de ressources matérielles est donc extrêmement délicat. Il n'existe pas aujourd'hui d'outils permettant ce partage de manière très efficace, simple et stable.

Mais justement, l'objectif de la grille recherche est de permettre l'expérimentation et le développement de nouvelles technologies de grilles. Une synergie entre grille de recherche et grilles de production existe et doit se développer. Des collaborations entre grilles régionales et grille de recherche permettent la valorisation des technologies de grille issues de la recherche menée dans Aladdin (on l'a vu pour CiGri mais aussi pour GRIF). De même, le partage d'expertise entre les ingénieurs et chercheurs impliqués dans la mise en œuvre, l'administration et le développement des grilles permet une meilleure diffusion des nouvelles méthodes et de nouveaux outils mais aussi l'identification de nouveaux besoins et de nouvelles limites. Un réseau d'experts, des séminaires réguliers communs entre IDG et Aladdin sur les limites des grilles actuelles et sur les prochaines technologies de grilles doivent être soutenus.

13 http://calcul.math.cnrs.fr/spip.php?rubrique6

14 http://www.cpu.fr/

15 www.genci.fr/

16 http://calcul.math.cnrs.fr/spip.php?article55

17 http://calcul.math.cnrs.fr/spip.php?article61

18 http://www.auvergrid.fr/

19 https://ciment.ujf-grenoble.fr/cigri

20 http://www.idgrilles.fr/

21 http://calcul.math.cnrs.fr/spip.php?rubrique19

22 http://calcul.math.cnrs.fr/spip.php?rubrique24

23 http://calcul.math.cnrs.fr/spip.php?rubrique30

24 https://ragrid.imag.fr/

25 http://graal.ens-lyon.fr/~diet/

26 http://www.eu-egee.org/

27 http://lappagenda.in2p3.fr/cdsagenda/fullAgenda.php?ida=a0774

28 http://calcul.math.cnrs.fr/spip.php?article6

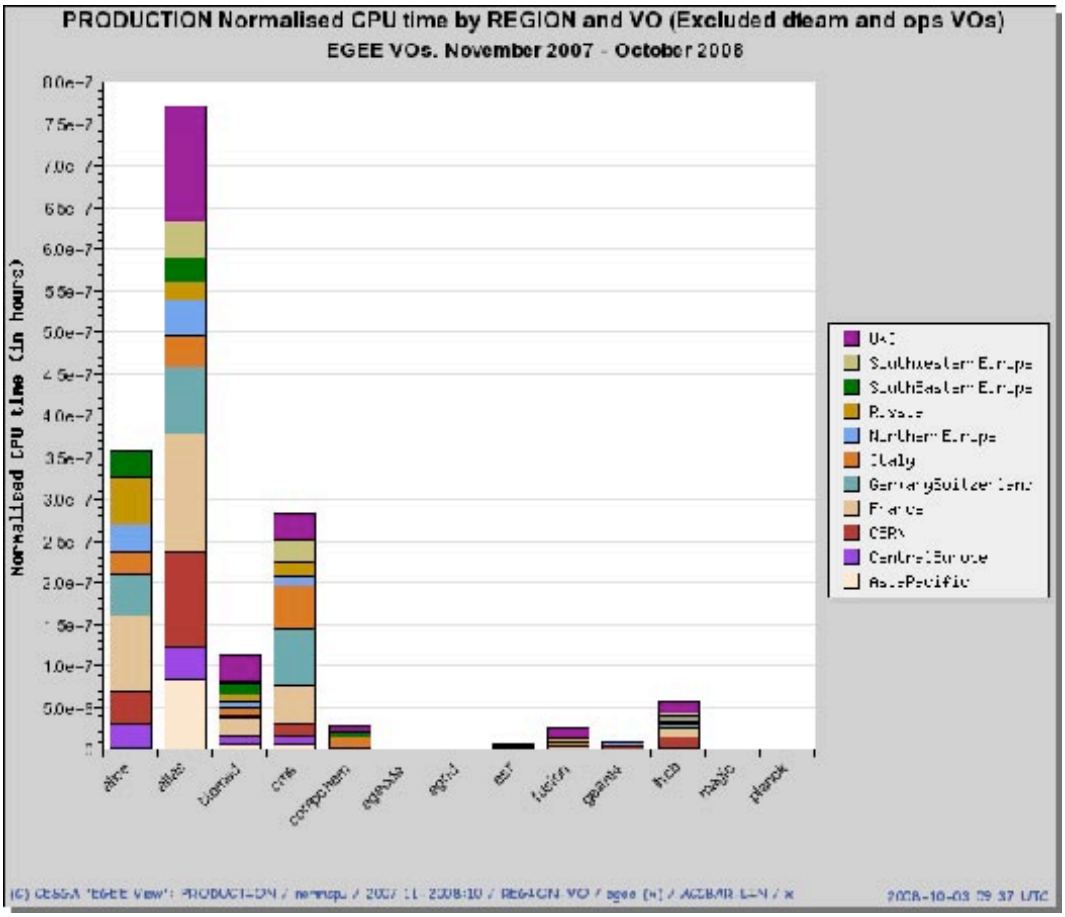
29 https://ciment.ujf-grenoble.fr/

30 http://www.flchp.univ-lyon1.fr/

Groupe transverse 3 : Relation avec les supercalculateurs

Cal Loomis, Victor Alessandrini, Serge Petiton, Danny Vandromme, Laurent Crouzet, Dominique Boutigny, François Robin, Philippe D'Anfray, Christine Rivière, Michel Kern

Les Grilles de calcul et les supercalculateurs sont à priori des mondes différents s'adressant à des communautés scientifiques distinctes. Les Grilles de calcul, dont le domaine de prédilection est celui de la physique des hautes énergies, sont bien adaptées pour des applications utilisant un grand nombre de processeurs sans communication entre ceux-ci (calculs paramétriques, Monte-Carlo, traitement de masses de données, etc.) alors que les supercalculateurs sont destinés aux modélisations nécessitant un haut degré de parallélisme et des communications intensives entre les processeurs. Toutefois il existe des passerelles entre ces deux mondes et certaines communautés commencent à utiliser les deux modes de calcul comme illustré par le graphique ci-dessous qui indique le temps CPU consommé par les principales Organisations Virtuelles de la grille EGEE sur une période d'une année. En dehors des grands utilisateurs que sont les expériences LHC, on voit apparaître d'autres disciplines telles que la biomédecine, la chimie, la fusion et les sciences de la terre.



1. Supercalculateurs et données

Il est possible de considérer les supercalculateurs comme des sources intenses de données au même titre que les grandes expériences telles que celles installées auprès du LHC. Par exemple dans le domaine de la simulation météo haute résolution, les problèmes traités aujourd'hui utilisent une centaine de matrices de 500 x 500 x 100 nombres de 64 bits qui subissent une évolution temporelle. Les résultats consistent en environ 150 sorties d'une vingtaine de Gigaoctets chacune.

Un projet complet produit 10 à 20 téraoctets qu'il faut stocker et rendre disponible pour la communauté des utilisateurs pendant 4 à 5 ans. En 2009, sur les nouvelles machines nationales il est prévu de doubler la taille

de la grille et de diviser le pas temporel par 2. La quantité de stockage correspondante va alors augmenter d'un facteur de l'ordre de 16. L'extrapolation des ressources nécessaires est complexe à réaliser, mais on voit qu'à terme, chaque projet de simulation météo haute résolution produira quelques pétaoctets de données qu'il faudra archiver et distribuer.

Cet exemple est à rapprocher des expériences sur le LHC[1] qui produiront jusqu'à 15 Pétaoctets de données chaque année. Ces données expérimentales ainsi que celles issues des simulations vont être distribuées et traitées sur une architecture de grille mondiale dans le cadre du projet W-LCG1. Il faut noter la grande compétence de la communauté de la physique des hautes énergies sur tous les aspects du stockage et de la distribution de masses de données scientifiques.

Dans un autre domaine, les calculs de Chromodynamique Quantique sur Réseau qui sont actuellement le facteur limitant pour l'exploitation de certains résultats de physique des particules, sont de très gros consommateurs de calcul. Un calcul s'exécute selon trois étapes:

• La première nécessite des supercalculateurs pour calculer des configurations de base, celles-ci sont stockées sur une grille de données (SRB ou Storage Resource Broker[2]) et mise à la disposition de la communauté mondiale,

• La deuxième phase (calcul des propagateurs) utilise des clusters de PC classiques ou une Grille de calcul et les résultats sont stockés sur une grille de données comme à l'étape précédente,

• La troisième étape (calcul des observables) nécessite une architecture classique (non parallèle) mais avec de grandes quantités de mémoire.

L'exemple de la chromodynamique quantique montre bien la complémentarité entre les différents moyens de calcul et surtout l'importance du stockage et de la distribution des données produites.

2. Les grilles de supercalculateurs

Dans le domaine des supercalculateurs, il faut mentionner l'existence de grilles telles que DEISA[3] en Europe ou Teragrid[4] aux États-Unis qui interconnectent chacune une dizaine de calculateurs. Toutefois, ces grilles ne servent pas à distribuer les calculs au niveau des applications elles-mêmes, elles servent plutôt à répartir la charge de manière intelligente entre les différents sites en partageant un espace de stockage global. En d'autres termes, chaque site est capable de faire tourner une instance différente d'une même application, mais une instance donnée est cantonnée à un seul site.

Des expériences ont été réalisées afin de faire tourner des applications parallèles et communicantes sur des grilles de calcul, mais le domaine d'application effectif est extrêmement restreint. En dehors de DEISA et de Teragrid qui concernent les très grands calculateurs, il est tout à fait possible d'intégrer le calcul parallèle dans des grilles de production telles que EGEE, à partir du moment où l'on considère les calculateurs parallèles comme des nœuds de la grille. Il faut pour cela être capable de définir correctement le nœud de calcul dans le système d'information de la grille et de faire en sorte que l'intergiciel soit capable de le prendre en compte. Des développements allant dans ce sens existent, ils se font, entre autre, dans le cadre des grilles régionales qui visent à intégrer les mésocentres.

De même, le projet WISDOM5 qui recherche des molécules efficaces dans le traitement des maladies négligées ou émergentes, utilise intensément les grilles afin de tester des molécules par des techniques de «docking». Des calculs de dynamique moléculaire requérant des machines parallèles sont nécessaires dans une deuxième phase. Il serait donc intéressant de pouvoir combiner les deux étapes sur une architecture de grille unique. Un projet allant dans ce sens est en cours de réalisation entre le centre de calcul KISTI à Daejon en Corée du Sud et le Centre de Calcul de l'IN2P3 dans le cadre du Laboratoire International Associé Franco-Coréen FKPPL6.

3. Relations avec les réseaux

L'ensemble du réseau informatique français est opéré par le GIP Renater[7]. Celui-ci est en train de déployer l'architecture Renater 5 (voir graphique ci-contre) qui est largement basé sur un maillage de fibres noires, c'est-à-dire de fibres optiques brutes et «allumées» par des équipements gérés par Renater.

La technologie optique DWDM de multiplexage en longueur d'onde permet de déployer avec une très grande souplesse, des liaisons à très haut débit (n x 10 Gb/s) entre les sites. Le déplacement de grandes quantités de données entre des grands centres informatiques n'est plus un problème à l'heure actuelle et ce, même à l'échelle internationale, grâce au réseau européen GÉANT qui est interconnecté avec le restant du réseau mondial. La souplesse de Renater 5 permet de créer des architectures ad hoc supportant les échanges de données des projets scientifiques. C'est ainsi que la partie française de la grille pour le LHC (LCG-France) peut s'appuyer sur un réseau performant.

Dans ce contexte, les futurs grands instruments, et en particulier les supercalculateurs, pourront compter sur le réseau Renater pour distribuer leurs données vers les mésocentres ou les grilles de production où elles pourront être post-traitées, archivées ou re-distribuées vers les utilisateurs finaux. La souplesse de Renater ne doit pas masquer la complexité des architectures sous-jacentes et il est important de planifier les développements longtemps en avance. Les nouveaux projets doivent donc dès le début intégrer une réflexion sur les besoins en termes de connectivité et de bande passante. Ceci est particulièrement vrai lorsque les utilisateurs sont très distribués.

4. Conclusions et recommandations

La communauté des grilles de production et, en particulier, celle liée à la physique des hautes énergies a développé un système sophistiqué et performant de stockage et de distribution des masses de données scientifiques. Parallèlement, les utilisateurs de supercalculateurs pointent la nécessité de développer considérablement les moyens de stockage en parallèle avec l'augmentation actuelle et à venir de la puissance de calcul. Un rapprochement entre ces communautés est donc indispensable autour de la thématique des données. L'utilisation des grilles de production pour la distribution des données produites par les grands calculateurs parallèles, renforcera l'interaction entre les supercalculateurs, les mésocentres et les grilles de production. Les données constituent donc un élément structurant pour la communauté du calcul intensif.

Concernant les réseaux qui forment l'ossature des grilles de production, il est indispensable de garantir la pérennité des financements et de maintenir une étroite collaboration entre les projets de calcul et Renater.

1 http://lcg.web.cern.ch/LCG/

2 http://www.sdsc.edu/srb/index.php/Main_Page

3 http://www.deisa.eu/

4 http://www.teragrid.org/

5 http://wisdom.eu-egee.fr/

6 http://fkppl.in2p3.fr/cgi-bin/twiki.source/bin/view/FKPPL/WebHome

7 http://www.renater.fr/

Groupe transverse 5 : Accès à la grille

Yonny Cardenas, Hélène Cordier, Virginie Dutruel, Rolf Rumler

L'objectif de ce rapport est d'estimer les coûts et les besoins en formation pour la France sur les quatre ans à venir. Il s'agit de dresser un bilan des utilisateurs et des actions pour faciliter les accès à la grille et ainsi bâtir une infrastructure de formation pérenne. Il nous faut dégager les grandes tendances aussi bien au niveau des besoins exprimés que des moyens utilisés. Ce rapport servira de base à l'élaboration d'un plan national coordonné par le ministère de la recherche.
Après une révision non exhaustive des ressources disponibles dans les plus grands projets grilles de production (EGEE et LCG), nous avons identifié les éléments suivants comme indispensables pour l'accès aux grilles.

1. FORMATION

Nous entendons par formation la période d'éducation durant laquelle le stagiaire acquiert des connaissances.

1.1. Bilan EGEE II France

Si nous dressons un bilan de la seconde phase de EGEE France, nous voyons que 24 formations d'une durée d'environ 2 ou 3 jours ont été organisées en France entre 2006 et 2008, dont 20 formations à destination des utilisateurs (dont 3 pour des domaines spécifiques : science de la terre, science de la vie), 3 pour les développeurs et 1 pour les administrateurs système.
Sur l'ensemble du projet EGEE et des 41 pays adhérents, 273 formations ont été organisées, soit 3334 participants. Nous pouvons donc constater l'importance de ces formations.
D'après le rapport d'activité NA3 (partie formation du projet EGEE) réalisé par Robin McConnell, le budget de l'activité NA3 du projet EGEE représente 5% du budget global de EGEE II.

1.2. Besoins exprimés, informations récoltées

À partir des autres rapports et des besoins qui ont été exprimés, il est évident et nécessaire de :
- Augmenter l'offre de formation et des supports. Beaucoup d'utilisateurs sont demandeurs de formations et cherchent à être opérationnels, il faut donc être réactif à leurs demandes.
- Former une large variété d'utilisateurs dans un maximum de domaines. C'est-à-dire qu'ils doivent avoir une formation sur l'utilisation générale des grilles de calcul mais également une formation sur une utilisation plus spécifique en fonction de leurs besoins, attentes et domaines de recherche.
- Développer des mécanismes de formation efficaces pour transmettre la connaissance aux utilisateurs. En effet, si un utilisateur suit une formation, pourquoi, dès son retour au sein de son équipe, il n'en ferait pas bénéficier ses collègues.
- Etendre et affiner le portefeuille d'outils pédagogiques en tenant compte de l'évolution des formations.
- Collaborer avec toutes les autres activités du projet afin de suivre les évolutions des différents domaines et ainsi toujours répondre à l'actualité et aux besoins qui en découlent. Les formations doivent être à jour et en aucun cas devenir obsolètes. Les utilisateurs sont les mieux placés pour faire part de leurs besoins et attentes spécifiques.

1.3. Infrastructure de formation

Les besoins étant exprimés et la nécessité des formations étant prouvée, à partir de la nouvelle organisation des grilles régionales, il serait judicieux d'organiser les formations autour des régions. Comment organiser une infrastructure de formation pérenne, l'objectif étant de stimuler et soutenir le développement des mécanismes de formation par régions émergentes.
- Avoir un contact formation par site (ou grille régionale) en relation directe avec une personne centrale qui coordonne les formations afin de faire remonter les besoins et d'avoir un interlocuteur privilégié au moment de l'organisation de formation.
- Proposer divers types de formations. Formations à destination des utilisateurs, des développeurs et des administrateurs mais il faudrait également mettre en place des formations par domaine scientifique afin de cibler le domaine de l'utilisateur, de répondre à ses besoins et demandes en respectant sa spécialisation. Il ne faut pas rester trop général mais essayer d'approprier au mieux la formation au public visé.
- Etablir un calendrier de formations afin de garder une bonne régularité et une bonne proportion entre les différentes formations proposées. Une régularité dans les formations permettrait une meilleure communication, diffusion de l'information et surtout organisation par rapport aux utilisateurs demandeurs de ces formations. Nous pourrions, par exemple, envisager une ou deux formations par région par an.
- Organiser des formations pour les formateurs afin de développer la base de formateurs. Il est important

d'élargir le nombre de ces experts formateurs mais il faut également améliorer l'expertise des experts et que chaque formateur soit préparé à enseigner en fonction de l'actualité des grilles de calcul. Ils doivent se tenir informé des nouveautés. Organiser des formations pour formateurs permettra de couvrir la diversification des formations.
Un répertoire de 89 formateurs a été créé lors de la seconde phase du projet EGEE. Nous établissons actuellement un répertoire pour la nouvelle phase de EGEE, que nous suivrons et actualiserons au mieux. Ce répertoire est essentiellement composé du personnel de certains sites EGEE ainsi que des utilisateurs experts de certaines VO (organisations virtuelles).
- Envisager de nous appuyer sur des entreprises ou associations d'experts, tel que Healthgrid, comme fournisseurs formation en termes de formateurs, infrastructures, et déplacements.
L'organisation de formations demande beaucoup d'efforts et de temps !
De nombreuses tâches administratives sont nécessaires:
- équipements, ressources, formateurs,
- communication,
- procédure d'inscriptions,
- logistique.

Ainsi que la préparation du matériel de formation:
- agenda,
- présentations,
- supports complémentaires,
- exercices pratiques.

Le but de la formation :
- Acquérir une autonomie et développer un soutien réciproque au sein de sa communauté,
- être capable d'utiliser les services de la grille,
- être capable de faire la formation à ses membres,
- améliorer la compréhension de ses propres applications.

2. ACCÈS A LA GRILLE

2.1. Structure grille pour la formation

De la même façon qu'il est nécessaire de mutualiser des ressources pour les grilles de production, il est nécessaire de mutualiser une légère proportion de ces ressources pour la formation en utilisant la même infrastructure.
La structure grille pour la formation doit comprendre essentiellement le déploiement et support d'une organisation virtuelle pour la formation (VO de Formation) dans les différents sites de production. Elle doit fournir une qualité de service convenable avec un temps de réponse raisonnable pour tourner des travaux dans le temps normal d'une session de formation.
Le déploiement de la VO «vo.formation.idgrilles.fr» sera proposé à l'ensemble des sites français de la grille de production EGEE/LCG. Il aura pour but d'offrir une infrastructure stable et pérenne aux différentes sessions de formation grille qui sont organisées en France.
Cette structure doit fournir en chaque site les ressources matérielles et humaines nécessaires pour la réalisation de différents types de formation (utilisateurs de base, développeurs, administrateurs). Cela implique la mise à disposition de ressources et services pour pouvoir faire les configurations dans le cadre de formation pour des administrateurs (similaire à un «testbed»).
Afin de se rapprocher le plus possible d'un environnement réel de grille de production, il est recommandé de déployer la VO de Formation sur plusieurs sites, composés de ressources et services hétérogènes. Elle doit permettre le déploiement des dernières versions du middleware de façon à anticiper sa mise en production.
Il est souhaitable de mettre en place une procédure spécifique rapide pour la demande de certificats. De même, la VO de Formation doit comprendre un portail consacré aux utilisateurs novices disposant de la plupart des fonctionnalités le plus avancées. Ces derniers servent de support à des tutoriaux et démonstrations poussées.
La VO de Formation assure, que ce soit aux utilisateurs ou aux administrateurs système, un accès à une première expérience sur les systèmes de grilles de calcul, permettant le transfert de connaissances et entre communautés scientifiques et l'industrie.

2.2. Structure pour le portage des applications

Une structure pour le portage des applications sur la grille est nécessaire. Elle doit être composée par une équipe d'ingénieurs qui assume le rôle de l'interface entre l'infrastructure grille et les utilisateurs de différentes disciplines scientifiques. Cette équipe doit participer au processus de « gridification » des applications et à l'assistance pour les nouveaux utilisateurs.

Dans ce contexte, cette structure doit participer à l'identification et à l'application de bonnes pratiques pour l'utilisation des outils et infrastructures disponibles. Ces expert doivent travailler avec les propriétaires de l'application afin d'établir la compréhension des besoins et l'identification des approches souhaitables. Pour cela, il est nécessaire d'établir et de mettre en œuvre une procédure pour la gridification des applications.
Dans le cadre du projet EGEE le «Grid Application Support Centre» (GASUC) fournit des services similaires. C'est une équipe composée de huit personnes : 5 développeurs, 1 web master, 1 directeur technique et 1 coordinateur de l'Institut de la Recherche en Informatique et Automatisation, en Hongrie. Ils ont des correspondants dans différentes institutions qui participent au projet EGEE. Ainsi le CEA est son correspondant en France.

2.3. Portail web pour l'exécution et surveillance des applications

Un portail web est également souhaitable pour permettre l'exécution et la surveillance des travaux dans différentes plateformes de grille. Cette ressource sera adressée aux utilisateurs novices. Le portail leur permet de gérer leurs travaux à travers une interface graphique conviviale. Il doit également laisser un degré de liberté appréciable aux utilisateurs avancés de la grille. Cet outil doit permettre la portabilité entre différentes plateformes de grilles avec un minimum de ré-ingénierie.
Dans le cadre de EGEE, le portail P-GRADE est proposé aux utilisateurs non-experts. Il cherche à dissimuler les mécanismes de bas niveau des architectures de la grille.

3. DISSÉMINATION

3.1. Site internet

Créer un site internet accessible à partir du site IDG http://idgrilles.fr afin de communiquer et diffuser au mieux les différents types de formations proposées, les formations programmées avec l'agenda de chacune d'elles et les supports et outils utilisés. Pour amener un maximum de personnes à se former, il faut les informer. Mais il faut également répondre aux mieux aux attentes des utilisateurs. Chacun doit pouvoir se renseigner directement et être au courant des dernières formations programmées.

3.2. Porte-feuille de supports de formation

Il est souhaitable de mettre en place un dépôt pour le partage et la réutilisation de supports de formation en relation avec les technologies grilles et son utilisation. Cet outil doit aider à agrémenter l'utilisation de la grille. Il s'adresse aux nouveaux utilisateurs qui peuvent trouver différentes ressources comme des articles, présentations, tutoriaux, etc… La mise à disposition à la communauté des utilisateurs, développeurs et administrateurs leur permettra d'ajouter du contenu approprié et des mises à jour.
Dans le projet EGEE il existe le « Digital Library » qui fournit ce service, mais il reste peu connu des utilisateurs.

4. ESTIMATION DES BESOINS EN FORMATION

Nous essayons d'estimer les besoins en formation sur les 4 ans à venir en nous appuyant sur les remarques et chiffres communiqués par les groupes thématiques.

Concernant le groupe « Planète - Univers » l'extension territoriale de la grille en des villes où les activités de ces disciplines sont importantes va amener de nouveaux utilisateurs ; il en est de même pour les grands projets quand ils seront dans leur phase de production. Pour l'instant il y a un formateur pour ces deux domaines, ce qui est insuffisant alors qu'il y a des besoins de formation régionale et nationale. Deux types de formation seront nécessaires, les unes pour les futurs experts, qui pourront ensuite participer à la formation, les autres pour de nouveaux utilisateurs. On peut estimer à 2 formations «thématiques» par an et à la participation de quelques futurs utilisateurs à d'autres formations proposées.

À ce jour, le potentiel offert par les grilles n'est pas encore identifié par les chercheurs des « Sciences Humaines et Sociales ». Il reste à convaincre la communauté. Une fois les chercheurs convaincus du potentiel de l'usage des grilles, nous pourrons évaluer les besoins en formation et à l'accès à la grille.

Concernant le groupe « Biologie – Santé », nous estimons approximativement à 1000 le nombre d'utilisateurs d'ici 4 ans, dont 200 sont déjà formés. Il nous faudrait donc former 800 utilisateurs, en recensant environ 10 formateurs dans ce domaine.

D'après le sondage effectué par le groupe « Chimie », la faible connaissance des grilles au sein de la communauté ne permet pas la mise en œuvre immédiate d'une démarche mettant en œuvre des actions de formation. Cette étape ne pourra être atteint que si un point de référence pour la Chimie a été créé et que des actions de sensibilisation et d'information sont menées.

Concernant le groupe « Science de l'Ingénieur et Informatique », nous recensons approximativement 100 utilisateurs. 65 utilisateurs seraient potentiellement intéressés par des formations et nous estimons à 8 le nombre de formateurs dans ce domaine.

La communauté des chercheurs en « Mathématiques – Physique – Fusion » ne semble pas très impliquée dans l'utilisation des grilles. Il est donc trop tôt pour évaluer les besoins en formation de cette communauté.

Concernant le groupe « Physique subatomique », nous recensons actuellement 477 utilisateurs potentiels de la grille. En 2012, il devrait y en avoir 565, il faudrait alors former 171 utilisateurs. Cela dit, parmi les utilisateurs potentiels de la grille, actuellement, les utilisateurs effectifs ne représentent pas plus de 30 %. Il faudrait donc pouvoir former de l'ordre de 245 utilisateurs sur 4 ans, soit 60 personnes par an.

En nous appuyant sur les chiffres d'EGEE et à titre de comparaison avec ce qui était écrit jusqu'ici à ce sujet, nous pouvons constater que le groupe « Hautes Énergies » est le plus important. En effet, il réuni 7582 utilisateurs, sur un total général de 11649 utilisateurs. Ce groupe représente donc 65 % du nombre total d'utilisateurs de la grille EGEE. Par déduction, leurs besoins représenteront donc 65 % des besoins globaux. Le domaine d'activité de ce groupe correspond essentiellement à celui du groupe « Physique subatomique » de l'Institut des Grilles. Une estimation des besoins totaux, incluant tous les groupes et basée sur les nombres d'EGEE donnerait 90 personnes à former par an pour l'IdG. On constate que le groupe « Biologie - Santé » tout seul est déjà plus grand et dépasse largement cette estimation par analogie. Il paraît que les utilisateurs potentiels de la « Grille française » viennent majoritairement d'autres disciplines que ceux d'EGEE. Une comparaison entre EGEE et l'IdG dans ce contexte est donc uniquement possible sans distinction par domaine scientifique.

Il reste à vérifier si l'évolution des mesures de formation dans EGEE peut servir comme base pour une estimation de celle de la formation pour l'IdG.

Groupes thématiques	Nombre d'utilisateurs potentiels d'ici 4 ans	Nombre de personnes à former potentiellement	Nombre de formateurs	Nombre de formations par an	Durée moyenne par formation
Science de l'ingénieur et nformatique	100	65	8	1	
Biologie, santé	1000	800	10	10	
Planète, univers, environnement	70 à 90	70 à 90	0	1 à 2	
Physique subatomique	565	245		4	3 - 4 jours

CONCLUSION

Ce livre blanc rend compte du travail intensif d'une dizaine de groupes de travail mis en place à l'occasion d'un exercice de prospective nationale sur l'intérêt scientifique des grilles de production. Les quatre questions sous-jacentes étaient les suivantes :

- Les grilles de production sont-elles une technologie nouvelle capable d'apporter aux chercheurs des différentes disciplines un outil décisif pour réaliser des percées scientifiques ? Dans quels domaines plus particulièrement ? Quels succès ont été obtenus et quels sont les objectifs à court et moyen terme de chaque communauté ?

- Quelle place doivent occuper les grilles de production dans l' « écosystème » des moyens de calcul offerts à la communauté de recherche française ?

- Quel investissement matériel est nécessaire pour répondre aux demandes émanant des différentes communautés ?

- Quels sont les besoins humains nécessaires et de quel type ?

Il ressort de la prospective une réponse très claire à la première question : oui, les grilles de production sont devenues en quelques années un outil indispensable à la communauté nationale dans plusieurs domaines scientifiques très importants : physique subatomique, sciences du vivant, sciences de la planète principalement. Les grilles de production apparaissent très clairement comme la ressource informatique, complémentaire des grands supercalculateurs, qu'il faut mettre à la disposition du plus grand nombre. Un grand effort de formation et d'information reste à accomplir car tous les domaines ne sont pas encore impliqués au même niveau et dans chaque domaine, tous les chercheurs ne sont pas informés ou formés au même niveau. Les 3 caractéristiques suivantes se dégagent cependant de l'ensemble des groupes thématiques :

- Il existe partout un noyau dur d'utilisateurs de la grille qui en sont très satisfaits et qui produisent des résultats scientifiques de haut niveau grâce à cette nouvelle technologie. La taille de ce noyau varie entre 10% et 60% des communautés sondées.

- La très grande majorité (environ 80%) des chercheurs encore peu impliqués dans les grilles jugent qu'il existe un très fort potentiel dans cet outil qu'ils souhaiteraient mettre à profit. Dans certains domaines scientifiques, une liste précise des projets envisagés est disponible.

- Il existe partout une grande part de la communauté encore très mal informée des grilles de production et des possibilités qu'elles pourraient apporter. Tout plan conséquent de déploiement doit donc obligatoirement s'accompagner d'un effort très substantiel de formation et d'information. Le point de blocage principal pour l'utilisation massive des grilles est le nombre insuffisant de « médiateurs », ingénieurs spécialisés possédant une bonne connaissance technique des grilles et une formation de base dans un domaine scientifique spécifique.

Les grilles de production doivent donc occuper une place reconnue et bien documentée au plus haut niveau dans l'écosystème des moyens de calcul. Les groupe de travail transverses « Grilles de données » et « Grilles et supercalculteurs » mis en place à l'occasion de cet exercice de prospective doivent poursuivre et amplifier leurs travaux pour encourager le développement de toutes les interfaces nécessaires au fonctionnement optimal de l'écosystème. De même, le rôle stratégique des grilles régionales a été souligné comme un pôle important de collaborations et de « capillarité ».

Une réponse précise concernant les besoins matériels n'a été fournie que par la communauté de physique subatomique. Celle-ci représente aujourd'hui 2/3 de l'utilisation totale des grilles mais ce chiffre pourrait descendre à 50% si les autres communautés amplifient leur utilisation. Le besoin total peut donc être estimé entre 1,5 et 2 fois les besoins exprimés par la physique subatomique, soit un besoin total de 100 kSI2k et 75 PétaOctets d'ici 2012, soit un besoin de financement d'environ 18,5 M€ sur 5 ans. Le plan d'investissement, décliné par site géographique, permettra de satisfaire ces besoins exprimés.

Les besoins humains recensés dans cette enquête correspondent à l'aspect information/formation /accompagnement des utilisateurs. Un total de 17 « médiateurs » a été recensé en sciences du vivant, sciences de la planète, sciences humaines et sociales, chimie et sciences de l'ingénieur et informatique pour les 5 prochaines années. Un renforcement de l'équipe centrale de formation est également nécessaire (2 FTEs). Ces postes ne sont pas nécessairement à pourvoir en personnel permanent. Ils ne prennent pas en compte les aspects opérationnels de la grille qui concerne la stabilisation du personnel impliqué dans la partie opérationnelle centrale de la grille française.

Domaine scientifique	**Nombre de médiateurs nécessaires dans les 2 ans à venir**
Sciences du vivant	5
Sciences de la Planète et de l'Univers	5
Sciences humaines et sociales	1
Chimie	2
Ingénierie et Informatique	4
Formation centrale	2
Total	19

En conclusion, ce livre blanc a permis de confirmer le grand potentiel scientifique des grilles pour la plupart des disciplines scientifiques, de mettre en évidence le grand intérêt qu'elles suscitent, et de quantifier l'ensemble des besoins matériels et humains nécessaires. Ces derniers apparaissent essentiellement sous la forme d'une vingtaine de médiateurs. Il apparaît par ailleurs extrêmement utile de voir se poursuivre les activités des groupes de travail transversaux mis en place lors de cet exercice de prospective. L'ensemble des recommandations des groupes de travail est compilé ci-dessous.

Groupe 1 : biologie-santé

Les recommandations s'appuient sur le travail des membres du groupe, notamment l'analyse des besoins et de l'état des lieux des grilles, ainsi que sur les résultats d'un sondage de la communauté réalisé au mois de mai 2008 au cours duquel plus de 400 réponses ont été collectées venant de plus de 60 laboratoires dans toute la France.

Les communautés de recherche dans le domaine de la biologie et de la santé identifient le besoin du déploiement d'une infrastructure de grille pour leur production scientifique. Une telle infrastructure permet de structurer et de fédérer les ressources informatiques de la communauté, de répondre aux besoins importants de calculs à la demande et offre une grande souplesse pour gérer dynamiquement et stocker les données distribuées et hétérogènes produites en volume croissant.

Complémentaire des supercalculateurs qui demeurent irremplaçables pour d'importantes applications hautement parallèles, la grille propose des modalités d'accès beaucoup moins contraignantes, une plus grande souplesse d'utilisation, la capacité de mutualiser des ressources et des services communautaires à coût réduit. La grille constitue ainsi un environnement original pour développer des collaborations et initier des recherches innovantes.

La communauté de recherche ne souhaite pas développer une grille indépendante pour les sciences du vivant et de la santé mais s'inscrit résolument comme partenaire de la mise en œuvre d'une ou plusieurs grilles pluridisciplinaires dans la mesure où celles-ci offrent toutes les garanties en termes de stabilité et de pérennité. La communauté peut exploiter aujourd'hui des grilles de production en France (Décrypthon, EGEE, grilles régionales) gérées par des intergiciels différents (Diet, gLite). Le fait que l'infrastructure nationale gère ses ressources à travers plusieurs intergiciels peut contribuer à améliorer la palette des services offerts aux utilisateurs mais cet enrichissement ne doit pas se faire au prix d'une instabilité ou d'une complexité accrue. L'utilisation systématique de standards internationaux (Open Grid Forum) dans la définition d'interfaces applicatives (Application Programming Interface) est vivement encouragée pour privilégier l'interopérabilité et la facilité d'utilisation. Le développement de services de haut niveau permettant de répondre aux besoins spécifiques des communautés d'utilisateurs en biologie et santé comme par exemple le Medical Data Manager sur EGEE, doit être poursuivi. L'existence d'une infrastructure (Grid5000) dédiée à la recherche en informatique et aux tests de nouveaux composants logiciels apporte de la flexibilité dans la validation d'approches innovantes mais il paraît essentiel pour cela de pouvoir y déployer tous les intergiciels (BOINC, Diet, gLIte) utilisés par les grilles de production et ainsi valider le passage à la grille d'applications dans les domaines de la biologie et de la santé avec ces intergiciels à la fois en termes de performances et de fonctionnalités.

Le besoin prioritaire unanimement identifié pour généraliser l'adoption des grilles est le recrutement d'ingénieurs, de bioinformaticiens et de neuroinformaticiens ou d'informaticiens dans le domaine de la santé qui servent d'intermédiaire entre les utilisateurs finaux et les équipes qui conçoivent les middlewares et administrent les sites de production. Ces personnels doivent être à même de comprendre les enjeux des projets scientifiques, d'analyser leur mise en œuvre sur la grille et de superviser leur déploiement. La distribution de ces ressources humaines doit être focalisée sur quelques équipes de recherche ou plates-formes jouant un rôle moteur et ayant déjà des compétences dans le but de former un réseau d'experts au service de l'ensemble de la communauté. Il parait aussi important de pouvoir regrouper une partie de ces moyens humains avec des plateformes de production de données et d'images en biologie et santé afin de mettre en place les services sur les grilles de données et de calcul au plus proche des utilisateurs finaux. Le démarrage de cette dynamique peut s'appuyer sur le recrutement d'une dizaine d'ingénieurs dont quelques-uns des ingénieurs expérimentés déjà présents dans la communauté sur postes permanents dès 2009 et 2010.

Les nœuds naturels d'une grille de production pour la biologie et la santé sont les nœuds actuels des grilles Décrypthon et EGEE ainsi que les plates-formes partenaires du réseau ReNaBi qui constituent dès à présent des pôles de compétences en bioinformatique pour la communauté. Pour la santé, la poursuite et l'extension du modèle actuel dans lequel les équipes de recherche s'appuient sur des grappes installées dans les universités (Décrypthon, grilles régionales) ou des centres de calculs (CC-IN2P3) permet de décharger celles-là de la charge d'exploitation d'un nœud de grille. L'accès des professionnels de santé à la grille depuis leur poste de travail à l'hôpital constitue une clef pour le développement de la recherche médicale. Cependant, les perspectives d'utilisation de la grille pour la routine clinique paraissent limitées à court terme, sauf dans certaines niches comme par exemple la planification de traitement en radiothérapie – curiethérapie. En revanche, ce besoin est

plus identifiable a court terme dans le domaine de la recherche clinique autour des plateformes de production d'images, y compris au sein des CHU équipés de telles plateformes de recherche.

Au niveau national, le Réseau National de Bioinformatique d'une part et le Groupement de Recherches STIC-Santé d'autre part apparaissent comme les meilleures structures où diffuser progressivement la culture des grilles et promouvoir leur utilisation par les communautés.

Au-delà, l'utilisation systématique de standards internationaux pour gérer l'interopérabilité des services et des bases de données permettra d'inscrire naturellement les ressources de la grille nationale dans les infrastructures de recherche européennes en cours de définition (FP7 design studies BBBMRI, EATRIS, ECRIN, EGI, ELIXIR, INFRAFRONTIER, INSTRUCT), renforçant ainsi la position des partenaires français impliqués dans ces projets.

Groupe 2 : Planète -Univers

- Nécessité du développement de la grille

Face à l'exploitation systématique des données archivées qui ont été peu exploitées jusqu'à présent, à de nouveaux instruments d'observation produisant toujours plus de données à traiter et à des simulations pour étudier plus finement les processus physiques, l'accès quotidien à des ressources informatiques importantes est une nécessité pour les équipes de recherche des Sciences de la Planète et de l'Univers. La Grille, plus que tout autre équipement informatique, est particulièrement adaptée à un certain nombre d'applications comme ce qui concerne l'exploration des espaces de paramètres ou le traitement massif de données. Le développement de la Grille pour offrir aux chercheurs les moyens informatiques qui leur permettront de répondre aux nouveaux enjeux scientifiques est une nécessité. Un des aspects important est aussi son potentiel pour la e-collaboration (partage de données, de résultats et d'outils).

Aspects opérationnels

Le nombre d'utilisateurs des Grilles de production peut exploser dans les prochaines années. Cependant, elle ne sera acceptée que si son développement se fait en prenant en compte les spécificités des communautés des Sciences de la Planète et des Sciences de l'Univers :

- elle doit supporter les environnements de travail les plus courants des communautés en terme de logiciels, langages, interfaces avec des sites de calcul et de données externes,
- elle doit être compatible avec la structuration de nos communautés en terme de centres de données et d'accès aux données dont les normes sont définies internationalement,
- son utilisation quotidienne ne se fera qu'à condition qu'elle soit stable et fiable dans le temps.

Extension géographique

Si la Grille intéresse les communautés Sciences de la Planète et Sciences de l'Univers, à part pour quelques équipes, son utilisation est encore modeste. Cette faible utilisation provient d'une méconnaissance de la Grille et est imputable au manque de contacts entre les experts de la Grille et les communautés Sciences de la Planète et Sciences de l'Univers, une des raisons, pas la seule, est qu'il n'y a pas toujours de sites de grille dans des villes-pôles de PU.

L'expertise grille se trouve naturellement là où sont localisés les points d'accès à la Grille. Afin que les communautés acquièrent le savoir-faire et qu'ils puissent l'utiliser au quotidien, il est nécessaire de profiter du développement de la Grille pour mettre en place un réseau territorial de spécialistes de la Grille dans les deux communautés. Il serait souhaitable que les principaux OSU disposent d'un nœud Grille avec du personnel formé.

Besoins d'information et de formation

L'organisation de réunion d'informations, la participation à des événements de communication comme la Ville Européenne des Sciences, et surtout la présentation de résultats scientifiques obtenus grâce à la Grille sont autant d'incitations à l'attention des scientifiques pour qu'ils prennent le temps de s'intéresser à ce nouveau mode de calcul.

Un verrou important est l'acquisition du savoir-faire par des informaticiens et scientifiques. L'organisation d'une formation continue est indispensable. Les formations doivent être dirigées vers :

- des experts locaux qui aideront ensuite les utilisateurs à porter leurs applications sur la Grille de façon à avoir un mode de fonctionnement comme il en existe dans les centres de calcul.
- les scientifiques pour leur apprendre à utiliser ce nouveau moyen de calcul et en découvrir les potentialités.

Le nombre de formateurs et de formations nécessaires est difficile à estimer au-delà du court terme. En effet, si la Grille répond aux attentes des communautés Sciences de la Planète et Sciences de l'Univers, le nombre d'utilisateurs peut exploser très rapidement. Il faudra tenir compte de l'organisation régionale et des demandes des participants qui auront à la fois besoin de tutoriel de base pour les nouveaux venus et de formation plus spécialisée et/ou appliquée.

La grille au sein de l'infosystème

Au sein des Sciences de la Planète et de l'Univers, les Grilles de production ne peuvent être qu'un moyen de calcul parmi d'autres. Les deux communautés continueront d'avoir besoin de super-calculateurs, de mésocentres, de moyens locaux et de Centres de Données hors de la Grille. La Grille a cependant toute sa place à prendre pour permettre aux scientifiques de traiter de nouveaux problèmes et de partager plus efficacement des données au sein de projets nationaux ou internationaux. Cela signifie qu'il est nécessaire de créer des interfaces pour passer de ces moyens de calcul à la Grille selon les besoins.

Les communautés Sciences de la Planète et Sciences de l'Univers se sont structurées autour de centres de données et ont développé des standards et des protocoles d'accès. Cette structuration et ces développements se sont fait hors de la Grille souvent au niveau international. Il est par conséquent nécessaire de réfléchir aux interfaces entre ces systèmes informatiques et les Grilles de production. Cette problématique a été / est abordée dans plusieurs projets européens (Open Geospatial Consortium, Global Earth Observation System of Systems, EuroVO-DCA, Open Grid Forum). Il est souhaitable que le développement d'une infrastructure Grille nationale prenne en compte ces spécificités et contribue à y apporter des solutions.

La poursuite des groupes de travail

Le groupe de travail « Sciences de la Planète et de l'Univers » mis en place dans le cadre de la prospective de l'Institut des Grilles a permis d'établir des collaborations entre les deux communautés. Afin de structurer les efforts dans le cadre de la gouvernance Grille pour ces communautés et de renforcer les collaborations naissantes, il semble judicieux de poursuivre l'existence du groupe de travail. Son rôle et sa structure restent à définir dans le cadre d'une NGI et devront être complémentaire des autres organisations (Actions spécifiques du CNRS, …) déjà en place sur cette thématique.

Aspects internationaux

Les projets sont de plus en plus internationaux. Par conséquent, la grille nationale doit être interopérable avec les grilles de nos partenaires, ce qui est le cas à l'heure actuelle avec EGEE. La structuration dans le cadre d'EGI qui sera mise en place devra tenir compte de ces collaborations internationales.

Groupe 7 : Physique subatomique

La communauté de physique subatomique bien historiquement pionnière sur les grilles de production devrait connaître dans les prochaines années, avec le démarrage du LHC et les études poussées de faisabilité de l'ILC, un accroissement continu des utilisateurs. D'autre part, le pouvoir attractif d'une grille de production arrivée à maturation et permettant de mutualiser facilement des ressources va dynamiser les utilisateurs des disciplines telles que l'astroparticule et la physique nucléaire. Globalement en 2012 nous escomptons que 85% des ressources seront utilisées au travers de la grille de production. Dans le même laps de temps la communauté (environ 1000 chercheurs) passera de 45% d'utilisateurs actuellement à 65% soit un accroissement de 10% par an.
Nous recommandons donc que l'effort entrepris jusque-là pour mettre sur pied une grille de production en France soit vigoureusement poursuivit avec une attention particulière pour les ressources dédiées au LHC afin de placer nos chercheurs dans une situation leur permettant de relever le défit de l'analyse des données du LHC dans un contexte de très forte compétition internationale. Les ressources LCG s'appuyant sur la grille EGEE, il est particulièrement important que la transition entre EGEE et EGI-NGI soit transparente aux utilisateurs.

D'autre part certains problèmes résiduels d'instabilité, de fiabilité et d'interopérabilité doivent être solutionnés dans le cadre du consortium gLite prévu par EGI. Il faudra également trouver une solution permettant d'utiliser des licences de certains produits payants tels que Mathematica par exemple sur les grilles de production.

Groupe 5 : Sciences de l'Ingénieur et Informatique

La situation de la communauté S2I est contrastée: une communauté informatique dans le domaine de la recherche et des applications sur les grilles de production existe; à l'inverse, il y a un déficit d'information considérable dans la communauté sciences de l'ingénieur, et aussi dans de nombreuses sous-communautés informatiques.
L'action de prospective a constitué par elle-même un outil de diffusion d'information, suscitant un intérêt et une attente qu'il importera de relayer: les deux communautés voient très majoritairement un intérêt à l'utilisation d'une grille de production. L'exploitation actuelle reste cependant modeste par rapport aux besoins réels. Pour se projeter dans ce modèle de production, la formation et le support doivent être adaptés. Mais surtout les spécificités suivantes du domaine S2I doivent être considérées.

- Du point de vue opérationnel, pour les sciences de l'ingénieur, le support des logiciels commerciaux de calcul scientifique, et au premier titre Matlab, est un prérequis. Le support à la diffusion et à l'exploitation des récents développements permettant l'accès à ces logiciels sur grille est une priorité.
- Du point de vue thématique, l'insertion de la grille dans l'écosystème, et les passerelles à établir avec la recherche sur les grilles concernent vivement S2I. Sur l'ensemble de ces questions, le rapport propose des pistes concrètes.
- Le rôle spécifique de la communauté S2I doit être mieux pris en compte. Son profil innovation/faible consommation de ressources, qui contraste avec celui d'autres communautés plus orientées vers l'exploitation, ne doit pas conduire à sous-estimer ses demandes.

Enfin, les décideurs et des utilisateurs ont besoin de lisibilité, pour identifier les échelles de temps et les coûts réels de mise en œuvre, et d'incitations à investir dans une technologie nouvelle. L'intégration de la thématique grille dans les programmes de l'ANR dont relèvent les communautés S2I y contribuerait significativement.

Groupe 4 : Chimie

La faible connaissance des grilles au sein de la communauté ne permet pas la mise en œuvre immédiate d'une démarche directe mettant en œuvre des actions d'information et de formation. Afin de mettre en place une action efficace, il apparaît tout d'abord nécessaire de créer un axe « Chimie » au sein d'une grille existante, puis d'initier une action de diffusion progressive sur son utilisation. Cette dernière sera effectuée en plusieurs étapes :

Création d'un point de référence pour la Chimie
Action de sensibilisation et d'information
Pérennisation du point de référence
Actions de formation (continue)

Création d'un point de référence pour la Chimie

Afin de répandre l'utilisation de grilles dans une communauté « vierge » comme celle de la Chimie, la meilleure stratégie consiste à agir en deux temps : une phase pilote suivie d'une phase d'extension ou phase opérationnelle. La phase pilote doit servir à créer un groupe de travail (chercheurs et techniciens) chargé de la création d'une grille de production en Chimie, délocalisée sur plusieurs sites nationaux ou à déployer des nœuds au sein d'une grille nationale déjà existante. Ce groupe de travail se chargera tout particulièrement d'identifier les problèmes liés aux applications numériques en Chimie, du portage de logiciels de Chimie sur la grille, du déploiement d'un nœud et de l'insertion dans le contexte européen (par exemple VOs Chimie dans EGEE). Cette démarche permettra l'identification d'un ensemble chercheurs-équipement-centres de recherches (et matériel hardware associé) spécialisé en Chimie et constituant un noyau de référence pour la création d'un réseau plus vaste qui sera déployé dans la seconde phase d'extension.

Une fois ces objectifs atteints, le groupe se focalisera sur des applications « phares » de la grille de production en Chimie, afin de bien montrer l'apport de cette technologie à la résolution de problèmes chimiques. La durée nécessaire à la réalisation des objectifs de la phase pilote peut d'ores et déjà être estimée à 24 mois environ.

La phase d'extension commencera une fois que la phase pilote aura pleinement atteint les objectifs escomptés, et sera progressivement mise en place. Elle devra donc répondre aux besoins exprimés et tenir compte des contraintes du moment.

Gouvernance

Le groupe qui a travaillé sur cette prospective regroupe désormais un nombre significatif de personnes représentatives scientifiquement ou géographiquement de la communauté française des théoriciens en Chimie. Il nous semble donc naturel que ce même groupe soit responsable de la gouvernance pour la partie Chimie et qu'il constitue le noyau du groupe de travail chargé de la première phase du projet.

Information et formation

Comme mentionné précédemment, l'activité du groupe de travail doit mettre en œuvre une politique globale de formation destinée à la fois aux gestionnaires du système présents dans chaque unité participante (labo ou équipe) et aux utilisateurs. En particulier, lors de la phase pilote une action de sensibilisation à l'usage des grilles (grâce aux exemples fournis par le groupe de travail sur les applications spécifiques en Chimie) sera effectuée au niveau national dans le cadre de séminaires dans des laboratoires et de conférences dans les congrès ou colloques.

La phase pilote se terminera par une première formation nationale « Grille de production en Chimie », centrée sur l'utilisation de grilles de production avec des exemples spécifiques en Chimie. Cette formation sera éventuellement reconduite comme formation continue destinée aux scientifiques et aux ingénieurs des laboratoires de Chimie (1 formation par an).

Moyens

La diffusion de l'utilisation de grilles en Chimie ne peut pas être limitée à des opérations de formation de chercheurs et techniciens déjà présents au sein des laboratoires. En effet, le succès de cette diffusion dépendra fortement de l'existence d'un personnel hautement spécialisé dans les applications des grilles en Chimie, permettant ainsi un appui logistique permanent. Dans la première phase du projet (24 mois), deux ingénieurs en CDD, chargés de monter et de gérer le projet pilote, pourront être prévus. Ceux-ci auront également en charge la liaison entre les différents laboratoires et équipes impliqués dans le projet. Dans la phase opérationnelle, une configuration minimale de l'équipe sera de deux ingénieurs et d'un chercheur permanent.

Dans le même temps, il faudra prévoir un financement destiné au déploiement d'au moins un nœud de grille de façon à disposer d'un point de référence « hardware » sur lequel les utilisateurs pourront accéder à la grille.

Enfin, des frais de missions en France et à l'étranger (Europe) seront à prévoir pour la participation à des réunions prévues dans le cadre du déploiement de la grille et à son intégration dans EGEE par exemple.

Groupe 3 : Sciences Humaines et Sociales

L'intégration de l'usage des Grilles pour les SHS représente une opportunité de structuration pour toutes les disciplines qui les composent. Tout est mûr pour une telle aventure mais il reste à convaincre la communauté de la chance qui se présente à elle de sortir des solutions numériques individuelles et non pérennes.

L'action à mettre en place peut se décomposer en quatre phases :

1) Aide à la structuration des acteurs du numérique en SHS

Aider au renforcement des Centres de Ressources Numériques qui se structurent au sein du TGE Adonis. Ils joueront le rôle d'interface entre les porteurs de projets de recherche et l'Institut des Grilles et ses opérateurs. Ainsi chaque CRN connaissant les possibilités opérationnelles ou les développements possibles saura pendre en charge l'aide au portage des données numériques sur les Grilles. Ils garantiront l'interopérabilité des données et leur pérennité au regard de la recherche. Cette action aura comme résultante de dériver progressivement les données numériques des chercheurs SHS au sein de structures informatiques moins disparates et dans des réseaux de haut de gamme et non plus dans des solutions souvent système « D » disparaissant avec les départs à la retraite. La recherche en SHS trouvera en cela une plus grande cohérence et la mise en place effective de silos de « connaissances » spécifiques à la discipline. Les financements de l'État trouveront dans cette démarche une meilleure visibilité sur la production de données « numérisées » et mutualisées. Cela limitera de fait la déperdition et facilitera l'interopérabilité.

2) Aide à l'émergence de projets « exemplaires »

Aider deux à trois projets pouvant servir d'exemples pédagogiques sur le potentiel de l'usage des Grilles. Des projets de recherche s'appuyant sur des base de données images de très haute définition, des calculs de films en images de synthèse HD, sur de la manipulation intensive de flux vidéo ou d'enregistrements sonores peuvent donner naissance à des cas d'école pour une meilleure prise de conscience du potentiel des Grilles.

3) Aide à la diffusion de ces pratiques

L'organisation d'une école thématique autour de ces usages pourra se faire une fois les premiers exemples validés. La diffusion de cahier de recommandations aux équipes de recherches fait partie de la sensibilisation de la communauté.

4) Aide par la mise en place d'appel à projets en SHS faisant l'usage des Grilles

Cette étape ne peut être atteinte que dans la mesure où les étapes antérieures l'ont également été.

RECOMMANDATIONS DES GROUPES TRANSVERSES

Groupe transverse 1 : Grilles de données

Le point de départ, pour toutes ces applications, est l'évolution générale vers la production et le stockage des données de façon distribuée, parfois près des lieux de production, mais pas toujours. Certaines applications utiliseront la Grille stricto sensu, d'autres certains éléments de la Grille. Ces quelques exemples illustrent la diversité des besoins, parmi lesquels on voit apparaître en particulier :

L'accès aux données à partir de programmes exécutés sur la grille

La sécurisation des accès, impérative dans certaines disciplines, et son corollaire, l'authentification des utilisateurs

La définition de standards pour assurer l'interopérabilité

Le développement d'outils – génériques, ou adaptés au contexte particulier d'une ou plusieurs applications - pour permettre la recherche, et l'indexation, dans un grand volume de données. Il s'agit d'un sujet de recherche en informatique, pour lesquels les interlocuteurs pertinents doivent être identifiés. Il faudrait certainement également encourager les échanges d'expériences dans ce domaine.

Des logiciels tels que le Storage Resource Broker (SRB) ou son successeur iRODS développés à San Diego (SDSC) sont des systèmes de grilles de données qui permettent de transférer et de cataloguer des données de manière simple. Il est également possible de définir des services associés aux données qui seront automatiquement appliqués en fonction de leurs provenances ou de leurs caractéristiques. SRB / iRODS est utilisé, entre autre, par la communauté biomédicale ; le réseau BIRN (Biomedical Informatics Research Network) aux USA est un exemple de déploiement à grande échelle d'une grille de données basée sur SRB.

Il faudrait, dans ce domaine en émergence, pérenniser ce groupe de réflexion. Il serait utile de mettre en place un point de rencontre régulier pour favoriser les interactions entre les équipes qui développent des applications et les spécialistes de la Grille, et partager les expériences dans les différents domaines. Il faudrait également disposer d'aide à l'implémentation et de tutoriels en ligne pour aider les débutants à s'approprier les technologies de la Grille.

Groupe transverses 2 : Grilles régionales

1- La synergie entre la « grille recherche » (GRID5000/Aladdin) et la « grille de production » est évidente. Les recherches développées au sein de la grille recherche permettent d'expérimenter de nouvelles technologies de grilles et donc de préparer les évolutions de la grille de production. Inversement, les besoins en matière de production peuvent induire des recherches spécifiques en matière de grille. Les contraintes et les objectifs d'utilisation de la grille recherche sont très différents de celles de la grille de production. Les techniques de virtualisation ne sont pas suffisamment mûres aujourd'hui pour une utilisation de la grille de recherche pour la

production. La séparation des deux architectures est nécessaire et utile. Les deux axes « grille recherche » et « grille de production » doivent donc être conjointement soutenus.

Il faut de même soutenir les groupes de travail communs « grille de production »/ »grille recherche » et les séminaires communs consacrés aux expériences, aux avancées et aux limites des technologies de grilles, afin de maintenir une cohésion forte entre les deux axes production et recherche.

2- Les mésocentres[31] ont pour vocation d'offrir des moyens de calcul intensif, de traitement et de stockage de données, des services associés et donc de développer l'expertise locale de la modélisation numérique et du calcul. Leur activité est principalement centrée sur la production locale et intermédiaire, l'accompagnement de projets scientifiques vers le calcul intensif et la formation. Ils servent de tremplin vers le centres de calcul nationaux. Cette activité est essentielle et doit être maintenue en particulier pour encourager les chercheurs à « penser PétaFlops ». Leur mise en réseau devrait être organisée au niveau national (demandée par la_CPU[32] avec le soutien de GENCI[33]) pour permettre une interaction structurée avec les centres de calcul nationaux d'une part et les grilles recherche et production d'autre part.

3- Les activités de grilles doivent se développer en lien avec les mésocentres (en particulier, dans ceux qui en expriment le besoin). Certes, certains problèmes de modélisations numériques ne relèvent pas des grilles de calcul mais, dans la mesure où le coût de traitement dans une grille devrait être moindre que dans un centre national pétaflopique, il faut encourager l'utilisation des grilles dans les mésocentres pour tous les problèmes qui en relèvent. Enfin et surtout, certaines communautés (climat, géophysique, astrophysique, biologie, etc.), partagent des données et des codes : la mise en grille de leurs mésocentres au niveau international (ou de leurs activités dans une grille de mésocentres) devrait renforcer le développement scientifique de ces communautés et devrait être soutenu.

3.1. Les mésocentres peuvent héberger des nœuds de grilles.

- Avantages : utilisation de l'infrastructure existante, permet de développer l'expertise des ingénieurs locaux et donc favoriser la diffusion des technologies de grille, dans toutes les communautés scientifiques.
- Limites : les ingénieurs des mésocentres sont déjà bien chargés => le coût de l'administration des nœuds locaux de grille doit être minimal sinon les équipes d'ingénieurs locaux doivent être renforcées. Il convient en particulier de mettre en œuvre des outils de grille standard dont l'administration doit être la plus légère possible.

3.2. Les utilisateurs doivent être formés à l'utilisation des grilles (introduction aux technologies de grilles, accompagnement de projets scientifiques à la mise en grille de leurs programmes). Il faut renforcer l'activité de formation et de soutien local aux projets scientifiques qui ont besoin des grilles. Les mésocentres pourraient être utilisées en relais de formation (cf. état des lieux des formations dispensées dans les mésocentres[34]). L'expérimentation des techniques desktop grids, ou toute autre technologie (cloud computing etc.), pour intégrer ponctuellement des ressources de mésocentres doit être soutenue.

4- Les grilles régionales devraient s'intégrer naturellement dans le paysage national ou international des grilles, soit en participant à une grille internationale de production (exemple de Grif[35] ou d'Auvergrid[36], soit en développant des liens étroits avec la grille recherche (exemple de CiGri[37]).

Groupe transverse 3 : Relation avec les supercalculateurs

La communauté des grilles de production et, en particulier celle liée à la physique des hautes énergies, a développé un système sophistiqué et performant de stockage et de distribution des masses de données scientifiques. Parallèlement, les utilisateurs de supercalculateurs pointent la nécessité de développer considérablement les moyens de stockage en parallèle avec l'augmentation actuelle et à venir de la puissance de calcul. Un rapprochement entre ces communautés est donc indispensable autour de la thématique des données.

L'utilisation des grilles de production pour la distribution des données produites par les grands calculateurs parallèles, renforcera l'interaction entre les supercalculateurs, les mésocentres et les grilles de production. Les données constituent donc un élément structurant pour la communauté du calcul intensif.

Concernant les réseaux qui forment l'ossature des grilles de production, il est indispensable de garantir la pérennité des financements et de maintenir une étroite collaboration entre les projets de calcul et Renater.

Groupe transverse 5 : Accès à la grille

Le nombre de formateurs et de formations nécessaires est difficile à estimer à court terme.
Il faudra tenir compte de l'organisation régionale et des demandes des utilisateurs qui auront besoin de tutoriaux de base pour les nouveaux venus et de formations plus spécialisées. Nous devons maintenir le groupe de travail et continuer à être attentifs aux différents besoins.
Nous nous attarderons sur la pérennisation de l'infrastructure de formation, notamment en suivant le déploiement de la VO 'vo.formation.idgrilles.fr' et la mise en place d'une procédure spécifique et rapide pour la demande de certificats.
Après mise en place d'un portefeuille de supports de formation, nous inciterons un maximum d'utilisateurs à se servir des ressources disponibles. Bien entendu, il faudra maintenir les outils de formation par rapport à l'évolution du middleware des grilles.
Enfin, nous souhaitons créer un site internet afin de communiquer et diffuser au mieux l'ensemble de l'activité 'Formation et accès à la grille' et nous cherchons un moyen d'interaction permanent avec les groupes thématiques et transverses.

[31] http://calcul.math.cnrs.fr/spip.php?rubrique6

[32] http://www.cpu.fr/

[33] http://www.genci.fr/

[34] http://calcul.math.cnrs.fr/spip.php?article55

[35] http://calcul.math.cnrs.fr/spip.php?article61

[36] http://www.auvergrid.fr/

[37] https://ciment.ujf-grenoble.fr/cigri

Considérant l'importance au plan national d'une coordination et de la mutualisation des compétences et des ressources, les Partenaires ont convenu de mettre en place une « démarche prospective nationale » organisée sur la période des neuf premiers mois de l'année 2008, permettant de réaliser : la sensibilisation des organismes et communautés à partir des savoirfaire et acquis, l'expression de leurs besoins, les synthèses intéressant les décideurs des Partenaires.

cf le Protocole d'accord concernant les « Grilles de Production » et la représentation française dans les projets européens de Grilles de Production

Entre les Partenaires:

L'Etat, à savoir le MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE

Le CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE,

Le COMMISSARIAT A L'ENERGIE ATOMIQUE,

L'INSTITUT NATIONAL DE LA RECHERCHE AGRONOMIQUE

L'INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE,

L'INSTITUT NATIONAL DE LA SANTE ET DE LA RECHERCHE MEDICALE,

La CONFERENCE DES PRESIDENTS D'UNIVERSITE,

Le GROUPEMENT D'INTERET PUBLIC POUR LE RESEAU NATIONAL DE TELECOMMUNICATIONS POUR LA TECHNOLOGIE, L'ENSEIGNEMENT ET LA RECHERCHE

RÉSUMÉ

Depuis 10 ans, les grilles informatiques tournées vers la production, font l'objet d'une activité de recherche et développement très intense en Europe. La communauté de physique des particules a joué un rôle pionnier et moteur dans l'établissement d'une véritable infrastructure de production opérationnelle pluridisciplinaire, qui est prête, notamment, à relever l'énorme défi posé par l'analyse des données du Large Hadron Collider (LHC) au CERN. D'autres communautés de recherche, notamment en sciences du vivant, de la planète et de l'univers, se sont intéressées très tôt à l'utilisation des grilles informatiques et ont démontré leur intérêt pour le traitement de grands volumes de données distribuées.

Ce livre blanc s'inscrit dans le contexte d'un exercice de prospective nationale sur l'intérêt scientifique des grilles de production. Les grilles de production apparaissent très clairement comme un outil nécessaire à de nombreuses disciplines pour de nouvelles avancées scientifiques, et comme très complémentaires des supercalculateurs. Ce livre fait le bilan de cet exercice de prospective, discipline par discipline, à partir d'études et de sondages réalisés en 2008, et met en avant des recommandations spécifiques tant pour chaque domaine scientifique que pour un certain nombre de thèmes transverses.